

日本国特許庁  
JAPAN PATENT OFFICE

04.06.2004

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日 Date of Application: 2004年 1月14日

出願番号 Application Number: 特願2004-006630

[ST. 10/C]: [JP2004-006630]

出願人 Applicant(s): 株式会社ポストゲノム研究所

REC'D 22 JUL 2004

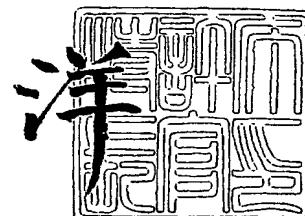
WIPO PCT

**PRIORITY DOCUMENT**  
SUBMITTED OR TRANSMITTED IN  
COMPLIANCE WITH  
RULE 17.1(a) OR (b)

2004年 7月 9日

特許庁長官  
Commissioner,  
Japan Patent Office

小川



BEST AVAILABLE COPY

出証番号 出証特2004-3059609

【書類名】 特許願  
【整理番号】 PGIA0301Y1  
【提出日】 平成16年 1月14日  
【あて先】 特許庁長官殿  
【国際特許分類】 C12Q 1/68  
【発明者】  
  【住所又は居所】 東京都台東区谷中 2-6-4 3-202  
  【氏名】 橋本 真一  
【発明者】  
  【住所又は居所】 千葉県松戸市松戸 159-1 第3住宅2-905  
  【氏名】 松島 紩治  
【発明者】  
  【住所又は居所】 東京都杉並区南荻窪 4-8-13  
  【氏名】 菅野 純夫  
【特許出願人】  
  【識別番号】 501005184  
  【氏名又は名称】 株式会社ポストゲノム研究所  
【代理人】  
  【識別番号】 100102978  
  【弁理士】  
  【氏名又は名称】 清水 初志  
【選任した代理人】  
  【識別番号】 100108774  
  【弁理士】  
  【氏名又は名称】 橋本 一憲  
【先の出願に基づく優先権主張】  
  【出願番号】 特願2003-402306  
  【出願日】 平成15年12月 1日  
【手数料の表示】  
  【予納台帳番号】 041092  
  【納付金額】 21,000円  
【提出物件の目録】  
  【物件名】 特許請求の範囲 1  
  【物件名】 明細書 1  
  【物件名】 図面 1  
  【物件名】 要約書 1

## 【書類名】特許請求の範囲

## 【請求項1】

次の工程を含む真核細胞の遺伝子タグの製造方法。

- (1) RNAのCAP部位にII<sub>s</sub>型制限酵素の認識配列を含むRNAリンカーを連結する工程、
- (2) (1)のRNAを錠型としてcDNAを合成する工程、
- (3) (2)のcDNAにRNAリンカーに含まれる認識配列を認識するII<sub>s</sub>型制限酵素を作用させ、遺伝子タグを生成する工程

## 【請求項2】

次の工程によってcDNAを合成する請求項1に記載の方法。

- i) RNAの任意の領域にアニールするプライマーによってcDNAの第1鎖を合成する工程、および
- ii) 第1鎖のRNAリンカーを錠型として合成された領域にアニールするプライマーによって、cDNAの第2鎖を合成して2本鎖cDNAとする工程

## 【請求項3】

第1鎖のRNAリンカーを錠型として合成された領域にアニールするプライマーが、固相に結合することができる標識を有するか、または固相に固定化されており、前記固相の回収によって2本鎖cDNAを回収する工程を含む請求項2に記載の方法。

## 【請求項4】

II<sub>s</sub>型制限酵素を作用させる前、または後に固相を回収する請求項3に記載の方法。

## 【請求項5】

RNAリンカーがII型制限酵素の認識配列を含む請求項1に記載の方法。

## 【請求項6】

遺伝子タグのII<sub>s</sub>型制限酵素の切断部位を、他の遺伝子タグのII<sub>s</sub>型制限酵素の切断部位と連結させて、ダイタグを生成する工程を含む、請求項1に記載の方法。

## 【請求項7】

RNAリンカーにアニールするプライマーによって、ダイタグを増幅する工程を含む請求項6に記載の方法。

## 【請求項8】

遺伝子タグのII<sub>s</sub>型制限酵素の切断部位に任意の塩基配列を有するアダプターを連結し、RNAリンカーと、前記アダプターにアニールするプライマーによって、遺伝子タグを増幅する工程を含む、請求項1に記載の方法。

## 【請求項9】

請求項1に記載の方法によって生成された遺伝子タグの複数を連結する工程を含む、遺伝子タグのコンカテマーの製造方法。

## 【請求項10】

請求項6に記載の方法によって生成されたダイタグの複数を連結する工程を含む、遺伝子タグのコンカテマーの製造方法。

## 【請求項11】

請求項9または請求項10に記載のコンカテマーの塩基配列を決定する工程を含む、遺伝子タグの塩基配列の決定方法。

## 【請求項12】

次の要素を含む、遺伝子タグの製造用試薬キット。

- (a) II<sub>s</sub>型制限酵素の認識配列を含むオリゴヌクレオチドからなるRNAリンカー
- (b) RNAリンカーをRNAのCAP部位に連結するための試薬
- (c) RNAリンカーを錠型として合成されたcDNAにアニールするオリゴヌクレオチドからなるcDNA第2鎖合成用のプライマー
- (d) cDNA第1鎖合成用プライマー

## 【請求項13】

cDNA第1鎖合成用プライマーが、以下のi)-iii)からなる群から選択されるいづれかのプライマーである請求項12に記載のキット。

- i) ランダムプライマー、
- ii) オリゴdTプライマー、および
- iii) 特定のmRNAに相補的な塩基配列を含むプライマー

【請求項 14】

次の工程を含む、真核細胞における遺伝子の発現プロファイルの取得方法。

- (1) 請求項 1 に記載の方法によって遺伝子タグを製造する工程、
- (2) (1)の遺伝子タグの塩基配列を決定する工程、および
- (3) 決定された塩基配列とその出現頻度を対応付けることによって発現プロファイルを得る工程

【請求項 15】

請求項 14 に記載の方法によって取得された遺伝子発現プロファイル情報を蓄積した、遺伝子発現プロファイルのデータベース。

【請求項 16】

請求項 14 に記載の方法によって異なる種類の細胞の遺伝子発現プロファイルを取得し、遺伝子発現プロファイルを比較して細胞間で発現頻度の異なる遺伝子タグを選択する工程を含む、遺伝子発現プロファイルの解析方法。

【請求項 17】

次の工程を含む、遺伝子の転写開始点の決定方法。

- (1) 請求項 1 に記載の方法によって遺伝子タグを製造する工程、
- (2) (1)の遺伝子タグの塩基配列を決定する工程、および
- (3) 決定された塩基配列をゲノムの塩基配列上にマッピングし、塩基配列が一致した領域を当該遺伝子の転写開始点として同定する工程

【請求項 18】

cDNAの第 1 鎮の合成のためのプライマーが特定の遺伝子の塩基配列から選択された塩基配列からなり、当該遺伝子の転写開始点を決定することを特徴とする請求項 17 に記載の方法。

【請求項 19】

次の工程によって決定された塩基配列またはその相補配列を含むcDNAを合成するための5'側のプライマーと、cDNAの任意の部位にアニールする3'側のプライマーを含む、cDNA合成用プライマーセット。

- (1) 請求項 1 に記載の方法によって遺伝子タグを製造する工程、および
- (2) (1)の遺伝子タグの塩基配列を決定する工程

【請求項 20】

3'側プライマーが下記の群から選択されたいずれかのプライマーである請求項 19 に記載のプライマーセット。

- i) オリゴdTプライマー
- ii) cDNAの断片配列情報、および
- iii) cDNAのII型制限酵素認識に隣接する遺伝子タグの塩基配列またはその相補配列からなるプライマー

【請求項 21】

次の工程を含む、全長cDNAの合成方法。

- a) 次の工程によって決定された塩基配列またはその相補配列を含むcDNAを合成するための5'側のプライマーと、オリゴdTプライマーからなる3'側のプライマーを用い、RNAまたはcDNAを錆型として相補鎖合成反応を行う工程、および

- (1) 請求項 1 に記載の方法によって遺伝子タグを製造する工程、および
- (2) (1)の遺伝子タグの塩基配列を決定する工程

- b) 合成されたDNAを全長cDNAとして回収する工程

【請求項 22】

請求項 21 に記載の方法によって得ることができる全長cDNA。

【請求項 23】

請求項22に記載の全長cDNAによってコードされるアミノ酸配列を含むポリペプチド。

【請求項24】

請求項23に記載のポリペプチドを認識する抗体。

【請求項25】

請求項22に記載の全長cDNAのコード領域を発現可能に保持するベクター。

【請求項26】

請求項25に記載のベクターを発現可能に保持する形質転換体。

【請求項27】

請求項26に記載の形質転換体を培養し、発現産物を回収する工程を含む、請求項23に記載のポリペプチドの製造方法。

【請求項28】

以下の工程を含む、請求項23に記載のポリペプチドの製造方法。

i) プロモーターに機能的に連結された請求項22に記載の全長cDNAのコード領域を含むDNAコンストラクトを、生体外翻訳を支持する要素と接触させる工程、および  
ii) 発現産物を回収する工程

【請求項29】

次の工程を含む、mRNAの5'末端の塩基配列を含むcDNAの合成方法。

a) 次の工程(1)-(2)によって決定された塩基配列またはその相補配列を含むcDNAを合成するための5'側のプライマーと、目的とするmRNAの任意の領域に対して相補的な塩基配列からなる3'側のプライマーを用い、RNAまたはcDNAを鑄型として相補鎖合成反応を行う工程、および

(1) 請求項1に記載の方法によって遺伝子タグを製造する工程、および

(2) (1)の遺伝子タグの塩基配列を決定する工程

b) 合成されたDNAをmRNAの5'末端の塩基配列を含むcDNAとして回収する工程

【請求項30】

請求項29に記載の方法によって回収されたcDNAの塩基配列を決定する工程を含む、mRNAの5'側の塩基配列を決定する方法。

【書類名】明細書

【発明の名称】遺伝子タグの取得方法

【技術分野】

【0001】

本発明は、遺伝子タグの取得方法、並びに遺伝子タグの解析方法に関する。

【背景技術】

【0002】

さまざまな細胞の遺伝子発現状態の比較によって、細胞を特徴付けることができる。つまり、細胞の状態を遺伝子の発現パターンで表現した細胞のカタログを得ることができる。このカタログを利用して、遺伝子の発現状態から細胞を特定することができる。逆に、細胞間で遺伝子の発現パターンを比較すると、各細胞に特徴的な遺伝子を拾い出すことができる。たとえば、正常な細胞と、人為的な処理を加えた細胞の間で遺伝子の発現状態を比較すると、人為的な処理を加えたときに発現レベルが変化する遺伝子が見出される。この遺伝子は、人為的な処理の結果として発現レベルが変化した遺伝子である。同様に、患者の細胞と健常者の細胞の間で遺伝子の発現状態を比較することによって、疾患に関連する遺伝子を見出すこともできる。

【0003】

このようにして、遺伝子の発現状態の比較によって、ある状態にある細胞で発現している遺伝子を網羅的に解析し、その種類や発現レベルを細胞間で比較することを、遺伝子の発現解析(expression analysis)と呼んでいる。遺伝子の発現解析のための手法には、さまざまな方法が用いられている。

【0004】

たとえば、以下に示す方法は、cDNAライブラリー間で発現レベルが変化している遺伝子を単離するために開発された方法である。

ディファレンシャルディスプレイ法(differential display)

サブトラクションライブラリー法(subtraction library)

【0005】

これらの方法は比較的古くから実用化されている方法である。いずれも由来の異なるcDNAライブラリーの間で、発現レベルの異なっている遺伝子を見出すための解析手法である。膨大な遺伝子の塩基配列情報が蓄積された近年においては、その塩基配列情報を利用した更に効率的な遺伝子発現解析が実現されている。すなわち、DNAアレイ法である。DNAアレイには、数万におよぶ遺伝子のプローブが高密度に配置されている。1つのDNAアレイを用いることで、一度の実験操作で数万の遺伝子の発現状態を知ることができる。ヒトの遺伝子の種類が3万～4万と推測されていることから、DNAアレイは、ヒトの遺伝子発現解析を強力に推進するツールとして広く普及しつつある。更にDNAアレイは、治療標的の探索や、薬剤候補化合物の開発に有用であると評価されている(Nature Genetics volume 32 supplement pp 547 - 552, 2002)。

【0006】

しかし、一般にDNAアレイを構成するプローブは、既知の塩基配列情報に基づいてデザインされている。したがって、未知の遺伝子の取得には不向きなデバイスである。更に、現在商業的に供給されているDNAアレイは、遺伝子配列情報が十分に蓄積された生物種に限られる。たとえばAffymetrics社は、次のような生物種についてDNAアレイを提供している。

シロイヌナズナ(Arabidopsis ATH1 Genome Array)

線虫(C. elegans Genome Array)

ショウジョウバエ(Drosophila Genome Array)

大腸菌(E. coli Antisense Genome Array)

ヒト(Human Genome Focus Array、他)

マウス(Mouse Expression Set 430、他)

緑膿菌(P. aeruginosa Genome Array)

ラット(Rat Expression Set 230、他)  
酵母(Yeast Genome S98 Array)

【0007】

DNAアレイによるその他の生物種の遺伝子発現解析のためには、スポットなどを利用しても利用者がDNAアレイを調製しなければならない。あるいは、カスタムアレイの作成サービスを利用する必要がある。それでも、遺伝子配列表の蓄積が不十分な生物種については、遺伝子配列情報に基づくDNAアレイを用意することは難しい。

【0008】

未知の遺伝子の取得を可能とし、しかも高度に効率的な遺伝子発現解析を可能とする手法として、SAGE(Serial analysis of gene expression)が提案された(SCIENCE, Vol. 270, 484-487, Oct. 20, 1995)。SAGEは、遺伝子に固有のタグを取得し、タグの塩基配列の塩基配列を網羅的に解析する解析手法である。遺伝子タグとは、その遺伝子の名札として利用することができる遺伝子の断片を言う。通常、10～20塩基程度の連続する塩基配列が、異なる遺伝子の間で完全に一致する可能性はそれほど高くない。たとえば9塩基からなる断片で、理論的には262144種類( $4^9$ )の遺伝子の識別が可能である。したがって、この程度の長さの断片は、遺伝子タグとして有用である。

【0009】

更にヒトゲノム配列において、18～21塩基からなるタグ配列の出現頻度と、そのタグ配列が遺伝子に固有の塩基配列である可能性は次のように計算される。

|    |                     |        |
|----|---------------------|--------|
| 18 | 268,435,456塩基に1回    | 89.43% |
| 19 | 1,073,741,824塩基に1回  | 97.24% |
| 20 | 4,294,967,296塩基に1回  | 99.3%  |
| 21 | 17,179,869,184塩基に1回 | 99.83% |

つまり、理論的には、18塩基のタグ配列では約90%以上、20塩基のタグ配列では約99%以上の確率で、遺伝子に固有の塩基配列であると考えることができる。ある遺伝子に固有の塩基配列は、遺伝子にユニークな塩基配列と呼ばれる。またゲノムにおいて、その出現頻度が1と見なされる塩基配列は、ゲノムにおいてユニークな塩基配列と呼ばれる。

【0010】

SAGEにおいては、II型制限酵素(Type IIIs Endonuclease)の作用を利用して、遺伝子タグが生成される。SAGEにおいてタグを生成するII型制限酵素は、タギング酵素と呼ばれる。II型の制限酵素がDNAの認識配列の中を切断するのに対して、IIIs型制限酵素は、認識配列から離れた位置を切断する。認識配列と切断位置の間の距離は、酵素によってほぼ一定である。たとえば、Bsm FIあるいはFokIは認識配列から9～10塩基の位置でDNAを切断し、粘着末端(sticky end)を残す。その他にも同様の作用を有するIIIs型の制限酵素として、次のような酵素が知られている(Szybalski, Gene 40:169, 1985)。

BbvI, BbvII, BlnI, FokI, HgaI, HphI  
MboII, MnII, SfaNI, TaqII, Tth111II

【0011】

更に、Mme Iと呼ばれるIIIs型制限酵素は、認識配列(5'-TCCRAC-3')から20塩基離れた位置を切断する(Tucholski et al, Gene Vol. 157, pp. 87-92, 1995)。MmeIをタギング酵素として利用し、20塩基長のタグを得ることができる発現解析方法も公知である(US Patent 6498013)。MmeIを利用するSAGEは、特にlong SAGEとも呼ばれる。以下に一般的なSAGEの原理を簡単にまとめた。

【0012】

まず、cDNAをII型制限酵素で切断し、その断片を回収する。II型制限酵素の認識配列が4塩基の場合、理論的には256塩基( $4^4$ )の断片に切断される。たとえばNla IIIの認識配列は4塩基である。cDNAの5'末端あるいは3'末端を固相に捕捉しておけば、切断されたcDNAの5'側、あるいは3'側の断片を、それぞれ容易に回収することができる。回収されたcDNAは2つの反応系に分割され、各反応系についてそれぞれ以下の操作が行われる。

## 【0013】

回収されたcDNAの切断箇所には、アダプターがライゲーションされる。アダプターは、末端にPCR増幅用のプライマーの塩基配列、中間にアンカーリング酵素の認識配列、そしてcDNAにライゲーションされる末端にII<sub>s</sub>型の制限酵素（タギング酵素）の認識配列が配置されている。2つの別のプールに分割されたcDNAには、それぞれに異なる塩基配列のプライマーの塩基配列を含むアダプターがライゲーションされる。アダプターのライゲーションの後にII<sub>s</sub>型の制限酵素を作用させると、II<sub>s</sub>型制限酵素はcDNAの末端を認識し、そこから離れた位置を切断する。こうして、II型制限酵素によって切断された部分から、II<sub>s</sub>型制限酵素に切断された部分までの断片からなるタグが生成される。生成されたタグは、ライゲーションされたアダプターを有している。

## 【0014】

II<sub>s</sub>型制限酵素の切断によって形成されたタグの粘着末端(sticky end)は、T4 DNAポリメラーゼによって平滑末端とされる。その後、前記の2つに分割された反応系のタグは、それぞれ平滑末端においてライゲーションされる。この結果、異なるプライマー配列を末端に配置して2つのタグが向かい合わせに連結される。2つのタグが連結されたものをダイタグと言う。ダイタグはPCRによって増幅され、アンカーリング酵素で切断される。その結果、PCRの増幅産物から、その両端のプライマー配列が除去される。更にプライマー配列を除かれたダイタグは、互いに連結されてダイタグのコンカテマーとする。こうして得られたコンカテマーがシーケンシングベクターに組み込まれる。

## 【0015】

コンカテマーの塩基配列を解析すれば、複数の遺伝子に由来する遺伝子タグの塩基配列を同時に明らかにすることができます。あるcDNAライプラリーから得られたコンカテマーの塩基配列情報を集積すると、理論的には、そのライプラリーを構成するcDNAの全ての遺伝子のタグ情報を得ることができます。こうして得られたタグ情報を、細胞間で比較すれば、容易に発現解析を行うことができる。

## 【0016】

DNAアレイによる発現解析には、塩基配列情報の蓄積が不可欠である。そのため、現在商業的に入手可能なDNAアレイは、ヒト、マウス、あるいは酵母などの一部の生物種に限られている。つまり、その他の多くの生物種において、DNAアレイを用いた遺伝子発現解析を行うためには、DNAアレイを新たに作成しなければならない。またDNAアレイは、既知の塩基配列情報に基づいて合成されたプローブ、あるいはクローニングされたcDNAをプローブとして用いる。その結果、一般的には、未知の遺伝子を見出すことは難しい。これに對してSAGEは、遺伝子の塩基配列情報の蓄積が不十分なことは、解析の障害とならない。更にプローブを必要としないSAGEは、未知の遺伝子の単離に有用な技術であると言える。

## 【0017】

しかし現在実用化されているSAGEのプロトコルにおいては、cDNAを制限酵素で切断し、得られた切断箇所にII<sub>s</sub>型制限酵素の認識配列を含むリンカーを連結している。したがって、SAGEに用いられる制限酵素には、認識配列が短いことが要求される。認識配列が長い制限酵素(rare cutter)では、切断されないcDNAが多くなってしまう。既知のSAGEにおいては、制限酵素で切断できないcDNAについては、タグが生成されない。

## 【0018】

たとえば4塩基を認識する制限酵素であるNlaIIIなどの制限酵素は、SAGEに好適であるとされている。理論的には、cDNAが4<sup>4</sup> (=256) 以上の長さを有していれば、NlaIIIの認識配列を少なくとも一つ含んでいいると言うことができる。確かに、256塩基以下の転写産物が存在する可能性は低いかもしれない。しかしライプラリーを構成する全てのcDNAが、常にNlaIIIの認識配列を含むとは限らない。つまり、256塩基以上の長さを有するcDNAであっても、タグが生成されない可能性はある。実際に、線虫の遺伝子をモデルとしたSAGEの評価において、NlaIII認識配列を持たないために、タグが生成されない遺伝子が存在することが報告されている(Genome Res. 2003 Jun. 13/6A:1203-15)。

## 【0019】

加えてこの工程を経て取得することができるタグは、cDNAを構成する塩基配列の中の制限酵素認識部位に隣接する塩基配列である。未知の遺伝子においては、cDNAの中のどこに制限酵素認識配列が存在するのかを予め予測することはできない。つまり、公知のSAGEによって取得されたタグの配列情報は、cDNAのどの場所に由来しているのかを予測することができないのである。

#### 【0020】

US Patent 6498013は、cDNAの5'側、あるいは3'側を捕捉することによって、それぞれ、5'側あるいは3'側のタグが得られることを示している。しかしこの工程によって生成されるタグは、cDNAの5'側あるいは3'側に位置する制限酵素(NlaIII)に隣接する塩基配列からなっている。言い換えれば、それは、cDNAに含まれるある制限酵素認識サイトのうち、もっとも5'側あるいは3'側にある制限酵素(NlaIII)に隣接する塩基配列である。つまり、cDNAの塩基配列のどこを占める塩基配列であるのかは、明らかでない。

遺伝子発現解析においては、タグを構成する塩基配列がcDNAの中のどこに由来しているのかは大きな問題とはならない。しかし、もしもタグの塩基配列がcDNAのどの部分を構成する塩基配列なのかを明らかにすれば、タグの有用性は更に高まる。

#### 【0021】

【非特許文献1】Nature Genetics volume 32 supplement pp 547 - 552, 2002

【非特許文献2】SCIENCE, Vol. 270, 484-487, Oct. 20, 1995

【非特許文献3】Szybalski, Gene 40:169, 1985

【非特許文献4】Tucholski et al, Gene Vol.157, pp.87-92, 1995

【非特許文献5】Genome Res. 2003 Jun. 13/6A:1203-15

【特許文献1】US Patent 6498013

#### 【発明の開示】

#### 【発明が解決しようとする課題】

#### 【0022】

本発明は、新規な原理に基づく遺伝子タグの取得方法、並びに遺伝子タグの解析方法の提供を課題とする。

#### 【課題を解決するための手段】

#### 【0023】

先に述べたように、現在実用化されているSAGEにおいては、制限酵素の認識配列に隣接する塩基配列がタグとして生成される。このことが、タグの塩基配列と、cDNAの全長配列との関係をわかりにくくさせていた。また、制限酵素の認識配列を含まないcDNAについては、タグが生成されないという課題を残していた。

本発明者は、制限酵素の認識配列に依存しないでタグを生成することができれば、これらの課題を解消できると考えた。たとえば、mRNAの5'末端を利用してタグを生成すれば、タグの塩基配列は様々な有用性を期待できるはずである。そこで、cDNAの合成方法として利用されていたCAP構造に着目し、遺伝子タグの取得への応用を試みた。その結果、mRNAの5'末端の塩基配列情報をタグとして取得できることを見出し、本発明を完成した。すなわち本発明は、以下のタグの取得方法、ならびにこの方法によって取得されたタグの用途に関する。

#### 〔1〕次の工程を含む真核細胞の遺伝子タグの製造方法。

- (1) RNAのCAP部位にII<sub>s</sub>型制限酵素の認識配列を含むRNAリンカーを連結する工程、
- (2) (1)のRNAを錆型としてcDNAを合成する工程、
- (3) (2)のcDNAにRNAリンカーに含まれる認識配列を認識するII<sub>s</sub>型制限酵素を作用させ、遺伝子タグを生成する工程

#### 〔2〕次の工程によってcDNAを合成する〔1〕に記載の方法。

- i) RNAの任意の領域にアニールするプライマーによってcDNAの第1鎖を合成する工程、および
- ii) 第1鎖のRNAリンカーを錆型として合成された領域にアニールするプライマーによって、cDNAの第2鎖を合成して2本鎖cDNAとする工程

〔3〕第1鎖のRNAリンカーを鋳型として合成された領域にアニールするプライマーが、固相に結合することができる標識を有するか、または固相に固定化されており、前記固相の回収によって2本鎖cDNAを回収する工程を含む〔2〕に記載の方法。

〔4〕II<sub>s</sub>型制限酵素を作用させる前、または後に固相を回収する〔3〕に記載の方法。

〔5〕RNAリンカーがII型制限酵素の認識配列を含む〔1〕に記載の方法。

〔6〕遺伝子タグのII<sub>s</sub>型制限酵素の切断部位を、他の遺伝子タグのII<sub>s</sub>型制限酵素の切断部位と連結させて、ダイタグを生成する工程を含む、〔1〕に記載の方法。

〔7〕RNAリンカーにアニールするプライマーによって、ダイタグを増幅する工程を含む

〔6〕に記載の方法。

〔8〕遺伝子タグのII<sub>s</sub>型制限酵素の切断部位に任意の塩基配列を有するアダプターを連結し、RNAリンカーと、前記アダプターにアニールするプライマーによって、遺伝子タグを増幅する工程を含む、〔1〕に記載の方法。

〔9〕〔1〕に記載の方法によって生成された遺伝子タグの複数を連結する工程を含む、遺伝子タグのコンカテマーの製造方法。

〔10〕〔6〕に記載の方法によって生成されたダイタグの複数を連結する工程を含む、遺伝子タグのコンカテマーの製造方法。

〔11〕〔9〕または〔10〕に記載のコンカテマーの塩基配列を決定する工程を含む、遺伝子タグの塩基配列の決定方法。

〔12〕次の要素を含む、遺伝子タグの製造用試薬キット。

(a)II<sub>s</sub>型制限酵素の認識配列を含むオリゴヌクレオチドからなるRNAリンカー

(b)RNAリンカーをRNAのCAP部位に連結するための試薬

(c)RNAリンカーを鋳型として合成されたcDNAにアニールするオリゴヌクレオチドからなるcDNA第2鎖合成用のプライマー

(d)cDNA第1鎖合成用プライマー

〔13〕cDNA第1鎖合成用プライマーが、以下のi)-iii)からなる群から選択されるいずれかのプライマーである〔12〕に記載のキット。

i)ランダムプライマー、

ii)オリゴdTプライマー、および

iii)特定のmRNAに相補的な塩基配列を含むプライマー

〔14〕次の工程を含む、真核細胞における遺伝子の発現プロファイルの取得方法。

(1)〔1〕に記載の方法によって遺伝子タグを製造する工程、

(2)(1)の遺伝子タグの塩基配列を決定する工程、および

(3)決定された塩基配列とその出現頻度を対応付けることによって発現プロファイルを得る工程

〔15〕〔14〕に記載の方法によって取得された遺伝子発現プロファイル情報を蓄積した、遺伝子発現プロファイルのデータベース。

〔16〕〔14〕に記載の方法によって異なる種類の細胞の遺伝子発現プロファイルを取得し、遺伝子発現プロファイルを比較して細胞間で発現頻度の異なる遺伝子タグを選択する工程を含む、遺伝子発現プロファイルの解析方法。

〔17〕次の工程を含む、遺伝子の転写開始点の決定方法。

(1)〔1〕に記載の方法によって遺伝子タグを製造する工程、

(2)(1)の遺伝子タグの塩基配列を決定する工程、および

(3)決定された塩基配列をゲノムの塩基配列上にマッピングし、塩基配列が一致した領域を当該遺伝子の転写開始点として同定する工程

〔18〕cDNAの第1鎖の合成のためのプライマーが特定の遺伝子の塩基配列から選択された塩基配列からなり、当該遺伝子の転写開始点を決定することを特徴とする〔17〕に記載の方法。

〔19〕次の工程によって決定された塩基配列またはその相補配列を含むcDNAを合成するための5'側のプライマーと、cDNAの任意の部位にアニールする3'側のプライマーを含む、cDNA合成用プライマーセット。

(1) [1] に記載の方法によって遺伝子タグを製造する工程、および  
 (2) (1)の遺伝子タグの塩基配列を決定する工程

[20] 3'側プライマーが下記の群から選択されたいずれかのプライマーである [19] に記載のプライマーセット。

i) オリゴdTプライマー  
 ii) cDNAの断片配列情報、および

iii) cDNAのII型制限酵素認識に隣接する遺伝子タグの塩基配列またはその相補配列からなるプライマー

[21] 次の工程を含む、全長cDNAの合成方法。

a) 次の工程によって決定された塩基配列またはその相補配列を含むcDNAを合成するための5'側のプライマーと、オリゴdTプライマーからなる3'側のプライマーを用い、RNAまたはcDNAを錠型として相補鎖合成反応を行う工程、および

(1) [1] に記載の方法によって遺伝子タグを製造する工程、および

(2) (1)の遺伝子タグの塩基配列を決定する工程

b) 合成されたDNAを全長cDNAとして回収する工程

[22] [21] に記載の方法によって得ることができる全長cDNA。

[23] [22] に記載の全長cDNAによってコードされるアミノ酸配列を含むポリペプチド。

[24] [23] に記載のポリペプチドを認識する抗体。

[25] [22] に記載の全長cDNAのコード領域を発現可能に保持するベクター。

[26] [25] に記載のベクターを発現可能に保持する形質転換体。

[27] [26] に記載の形質転換体を培養し、発現産物を回収する工程を含む、 [23] に記載のポリペプチドの製造方法。

[28] 以下の工程を含む、 [23] に記載のポリペプチドの製造方法。

i) プロモーターに機能的に連結された [22] に記載の全長cDNAのコード領域を含むDNAコンストラクトを、生体外翻訳を支持する要素と接触させる工程、および

ii) 発現産物を回収する工程

[29] 次の工程を含む、 mRNAの5'末端の塩基配列を含むcDNAの合成方法。

a) 次の工程(1)-(2)によって決定された塩基配列またはその相補配列を含むcDNAを合成するための5'側のプライマーと、目的とするmRNAの任意の領域に対して相補的な塩基配列からなる3'側のプライマーを用い、RNAまたはcDNAを錠型として相補鎖合成反応を行う工程、および

(1) [1] に記載の方法によって遺伝子タグを製造する工程、および

(2) (1)の遺伝子タグの塩基配列を決定する工程

b) 合成されたDNAをmRNAの5'末端の塩基配列を含むcDNAとして回収する工程

[30] [29] に記載の方法によって回収されたcDNAの塩基配列を決定する工程を含む、 mRNAの5'側の塩基配列を決定する方法。

#### 【発明の効果】

#### 【0024】

本発明は、 mRNAの5'末端の塩基配列を遺伝子タグとして取得する方法を提供した。mRNAの5'末端は、真核細胞のmRNAの全てが有する構造である。したがって、 mRNAの塩基配列にかかわらず、原理的に全ての遺伝子からタグを取得することができる。一方、公知の原理に基づくSAGEは、制限酵素認識サイトに隣接する領域をタグとして生成する。その結果、もしもmRNAを構成する塩基配列に制限酵素認識サイトが含まれなければ、その遺伝子のタグを取得することはできない。したがって、全ての遺伝子のタグを取得できる原理を提供した点において、本発明の意義は大きい。

#### 【0025】

また本発明のタグの方法によれば、 mRNAの断片からも遺伝子のタグを取得できる可能性がある。生体試料に含まれるRNAは、様々な原因によって、常に分解の危険にさらされている。したがって、cDNAの取得、あるいは得られたcDNAに基づく種々の解析結果は、 mRNA

の保存条件に大きく左右される。SAGE法も、mRNAの構造が完全に維持されていない場合には、遺伝子タグを取得できなかったり、あるいはタグの再現性が失われる可能性がある。

#### 【0026】

しかし本発明の方法によれば、mRNAの5'末端をタグとして取得することにより、たとえmRNAが断片化されていても、5'末端の構造さえ維持されていれば、正しくタグを取得することができる。したがって、mRNAの保存状態の影響を受けにくい。この特徴は、遺伝子の発現解析の信頼性を高める。

#### 【0027】

更に、本発明によって得ることができるタグの塩基配列は、mRNAの5'末端の塩基配列からなっている。その結果、本発明によって得られるタグの塩基配列情報は、さまざまな分野に応用することができる。たとえば以下のような用途は、本発明のタグによって始めて実現された用途である。

ゲノムにおける転写開始点の同定

全長cDNAの合成用プライマーの提供

cDNAライブラリーの全長率の評価

既知の原理に基づくSAGEによって得られたタグは、mRNAのどの領域の塩基配列なのかが明らかでない。したがって、このような用途に用いることはできない。

#### 【発明を実施するための最良の形態】

#### 【0028】

本発明は、次の工程を含む真核細胞の遺伝子タグの製造方法に関する。

- (1) RNAのCAP部位にII<sub>s</sub>型制限酵素の認識配列を含むRNAリンカーを連結する工程、
- (2) (1)のRNAを錫型としてcDNAを合成する工程、
- (3) (2)のcDNAにRNAリンカーに含まれる認識配列を認識するII<sub>s</sub>型制限酵素を作用させ、RNAの5'末端配列からなる遺伝子タグを生成する工程

#### 【0029】

CAP構造は、真核細胞あるいは真核細胞に感染するウイルスのmRNAの5'末端に存在する構造である。具体的には、7-メチルグアノシンが5'-5'-3'リン酸架橋を介してmRNAの5'末端のヌクレオチドに結合してCAP構造を構成している。mRNAはCAP構造によって5'-3'エクソヌクレアーゼ活性による分解から保護されている。細胞内では、役割を終えたmRNAのCAP構造は、デキャッピング酵素(decapping enzyme)によって除去される。その結果、CAP構造を失ったmRNAは、5'-3'エクソヌクレアーゼによって分解される(LaGradeur et al., EMBO J, 17:1487-1496, 1998)。CAP構造は、RNAポリメラーゼIIによる転写反応の初期の段階でRNAの5'末端に付加されていると考えられている。

#### 【0030】

本発明の方法は、このRNAのCAP構造にRNAリンカーを連結する工程を含む。本発明において、RNAは、真核細胞に由来するあらゆるRNAを用いることができる。より具体的には、polyA(+)RNAやtotal RNAを用いることができる。具体的には、動物、植物、酵母、あるいは粘菌などの、mRNAにCAP構造を有するあらゆる生物種に由来する細胞を利用することができます。

#### 【0031】

更に、これらの真核細胞に感染するウイルス由来のRNAも、CAP構造を有している。したがって本発明においては、真核細胞に由来する、真核細胞に感染、あるいは導入された遺伝子情報を転写したRNAも、真核細胞に由来するRNAに含まれる。真核細胞に感染した遺伝子の情報とは、たとえば、ウイルス、ウイロイド、あるいはマイコプラズマのような細胞内寄生体の遺伝子情報が含まれる。これらの遺伝子情報は天然のものであってもよいし、人為的に構成されたものであっても良い。一方、真核細胞に導入された遺伝子の情報とは、ベクターなどによって、人為的に導入された遺伝子情報を言う。たとえば、本来CAP構造を持たないとされている原核細胞の遺伝子であっても、転写可能な形で真核細胞に導入することによって、CAP構造を与えることができる。こうして転写されたRNAも、本発明における真核細胞に由来するRNAに含まれる。

## 【0032】

これらの細胞からRNAを抽出し、本発明の方法に利用する。RNAの抽出方法は公知である。市販のRNA抽出用のキットを利用すると便利である。たとえばRNAeasy(QIAGEN)などの市販のキットを利用して、高純度のRNAを容易に得ることができる。RNAの抽出にあたり、細胞の破壊が必要な場合には、公知の方法によって破壊することができる。

## 【0033】

本発明において、CAP構造に連結するRNAリンカーは、少なくともII<sub>S</sub>型制限酵素の認識配列を含むオリゴヌクレオチドからなる。RNAリンカーとして用いるオリゴヌクレオチドは、DNAであってもRNAであっても良い。好ましいRNAリンカーはRNAである。RNAリンカーを構成する塩基配列は、II<sub>S</sub>型制限酵素の認識配列を含む任意の塩基配列であってよい。ただしII<sub>S</sub>型制限酵素の認識配列は、RNAリンカーの3'末端に配置することが望ましい。

## 【0034】

II<sub>S</sub>型制限酵素は、その認識配列を基準として一定の塩基数だけ離れた位置を切断する。本発明は、mRNAの5'末端をタグとして取得することを目的としている。したがって、mRNAの5'末端にできるだけ近くに認識配列を配置することが望ましい。RNAリンカーを構成するII<sub>S</sub>型制限酵素の認識配列は、解析に用いるII<sub>S</sub>型制限酵素に合わせてデザインすることができる。たとえばMmeIの認識配列は5'-TCCRAC-3' (R=G or A)であることは既に述べた。したがってRNAリンカーは、その3'末端に、この塩基配列を配置するのが望ましい。なおII<sub>S</sub>型制限酵素の認識配列は、II<sub>S</sub>型制限酵素がその3'側を切断するように配置する。

## 【0035】

本発明のRNAリンカーとして有用な塩基配列を以下に示す。この塩基配列は、3'末端に配置されたII<sub>S</sub>型制限酵素(MmeI)の認識配列(TCCRAC;大文字)に加え、II型制限酵素であるXhoIの認識配列(cucgag;アンダーライン)も含んでいる。

5' -oligo 1 (配列番号：1) :

5'-uuuggauuugcuggugcaguacaacuaggcuuaauacucgagUCCGAC-3'

5' -oligo 2 (配列番号：2) :

5'-uuucugcucgaauucaaggcuuacgauguacgcucgagUCCGAC-3'

## 【0036】

付加されたXhoIサイトは、タグの連結、そしてベクターへの組み込みに利用することができる。更に、RNAリンカーを構成する塩基配列は、タグの増幅のためのプライマーがアニールするための領域として利用することもできる。プライマーがアニールするためには、アニーリングのための領域が、少なくとも15塩基、通常20-50塩基、たとえば20-30塩基で構成されるのが好ましい。またその構成塩基は、プライマーの融解温度(T<sub>m</sub>)が、通常60-80°C、たとえば65-75°C程度を有するようにデザインすることができる。プライマーがアニールするための塩基配列は任意である。したがって、たとえば、前記T<sub>m</sub>を与えることができる任意の塩基配列を利用することができる。

## 【0037】

プライマーがアニールするための塩基配列は任意である。更に、各種の制限酵素の認識配列を構成する領域と、プライマーをアニールさせるための領域は、RNAリンカーの中でも重複させることもできる。ただし、2種類のRNAリンカーに対して異なるプライマーをアニールさせる場合には、重複しないようにデザインすることによって、アニーリングの特異性の向上が期待できる。

## 【0038】

本発明において、RNAリンカーは、RNAのCAP構造に連結される。CAP構造にオリゴヌクレオチドを連結するための方法は任意である。たとえばオリゴキヤッピング法は、本発明におけるRNAリンカーの結合のための好ましい方法である。オリゴキヤッピング法は、mRNAの5'側の塩基配列を保持したcDNAを合成するためにによって開発された方法である(Maruyama, K and Sugano, S.:Gene 138: 171-174, 1994)。オリゴキヤッピング法においては、mRNAの3'末端poly(A)配列と、5'末端のCAP構造に連結されたRNAリンカーの塩基配列を利用して、全長cDNAの取得が実現されている。5'側の塩基配列が不完全なmRNAはCAP構造を保持

していないので、RNAリンカーが連結されない。そのため、オリゴキャッピング法においては、全長cDNAを特異的に取得することができた。

#### 【0039】

以下にオリゴキャッピング法の反応原理を簡単に述べる。まずmRNAをバクテリアアルカリ性フォスファターゼ(BAP)で処理して、CAP構造を持たないRNAの5'末端のリン酸基を加水分解する。この過程でCAP構造を備えていないRNAは、5'末端のリン酸基を失う。すなわち、断片化したRNA、あるいはミトコンドリア由来のRNAなどの5'末端に突出しているリン酸基が除去される。次いでタバコ酸性ピロフォスファターゼ(TAP)を作用させる。TAPはCAP構造のトリリン酸結合を加水分解する。その結果、CAP構造を有するRNAに特異的に5'末端のリン酸基を与えることができる。

#### 【0040】

BAPおよびTAP処理したRNAには、RNAリンカーが連結される。RNAリンカーの結合は、たとえばT4 RNAリガーゼを利用することができる。T4 RNAリガーゼによるライゲーションは5'末端のリン酸基を要求する。したがって、TAPによって5'末端リン酸基を得たRNAに対して特異的にRNAリンカーが連結される。こうして、CAP構造特異的にRNAリンカーを結合することができる。なおRNAを取り扱う反応においては、全ての工程をRNaseを排除された環境で行うことが望ましい。

#### 【0041】

オリゴキャッピング法には、いくつかのバリエーションが報告されている。たとえばCAP結合蛋白質カラムを利用して、CAP構造を有するRNAを精製する方法が知られている(Eddy, L. et al., Mol. Cell Biol. 15: 3363-3371, 1995)。この方法を利用すれば、CAP構造を有するRNAを固相上に捕捉することができる。固相を洗浄してCAP構造を有しないRNAを除去後にTAPで処理すれば、CAP構造を有しているRNAを回収することができる。こうして回収されたRNAは、5'末端にリン酸基を有するので、そのままRNAリンカーを連結することができる。すなわちCAP結合蛋白質を利用する方法は、BAP処理を必要としない。

#### 【0042】

次いで、RNAリンカーを連結したRNAを錆型としてcDNAが合成される。cDNAを合成するための方法は任意である。以下に、cDNAを合成するための方法について、代表的な方法を記載する。

一般にcDNAの合成は、第1鎖の合成と、第2鎖の合成の二つのステップで構成される。第1鎖の合成は、RNAを錆型として利用する逆転写反応である。これに対して第2鎖は、先に合成された第1鎖DNAを錆型とする相補鎖合成反応によって合成される。それぞれ、反応を開始するプライマーによって特徴付けられるいくつかの反応が知られている。

#### 【0043】

本発明において、cDNAの第1鎖は、RNAの任意の領域にアニールするプライマーによって合成することができる。RNAを錆型として逆転写酵素活性を利用してDNAを合成する方法は公知である。具体的にはMMLV由来の逆転写酵素(Reversetranscriptase;RT)やその変異体などを利用し、プライマーの伸長反応によって第1鎖を合成する方法が公知である。逆転写酵素の変異体としては、逆転写酵素が有するRNaseH活性を失わせた変異体(Superscript II, Gibco BRL)などが市販されている。またTth DNAポリメラーゼのように、DNA合成酵素でありながら、RNAを錆型とする相補鎖合成反応を触媒する酵素も知られている。このような酵素を利用すれば、第1鎖(RNA template)の第2鎖(DNA template)を単一の酵素で合成することもできる。続いてcDNAの合成のためのプライマーについて記載する。

#### 【0044】

先に述べたオリゴキャッピング法においては、通常、第1鎖の合成にはオリゴdTプライマーが利用される。cDNAの全長を合成するためには、第1鎖の3'末端から合成しなければならないため、mRNAの3'末端を占めるpoly(A)に相補的な塩基配列を有するオリゴdTプライマーが利用される。本発明においても同様に、オリゴdTプライマーを利用することによって、全長cDNAの5'末端をタグ配列として取得することができる。

#### 【0045】

これに対して本発明においては、RNAの全長は必ずしも必要ではない。本発明においては、タグはRNAの5'末端を含むわずかの領域から取得される。したがってRNAの5'末端を含む領域がcDNAとして合成できれば、本発明に必要なcDNAを得ることができる。したがって、たとえばRNAの任意の部分から相補鎖を開始できるランダムプライマーを利用して第1鎖を合成することができる。ランダムプライマーの利用によって、3'側の塩基配列が不完全な断片であっても、CAP構造を有するRNAであればタグを取得することができる。ランダムプライマーは、より幅広いRNAからタグを取得できる点で、特に遺伝子発現解析においては有用なプライマーである。

#### 【0046】

更に、第1鎖の合成にあたって、特定の遺伝子の塩基配列に相補的な塩基配列を有するプライマーを利用することによって、特定の遺伝子のタグを選択的に取得することもできる。たとえば、部分的な塩基配列のみが明らかにされ、5'側の塩基配列が不明な遺伝子について、本発明を利用して5'末端のタグ配列を取得することができる。そのためには、第1鎖の合成に当たって、明らかにされている塩基配列からプライマーとする塩基配列を選択する。このプライマーは、mRNAの明らかにされている領域から5'末端にかけての領域をcDNAの第1鎖として生成する。プライマーは特定の遺伝子の塩基配列から選択されているので、目的とする遺伝子以外のRNAからは第1鎖が生成されない。その結果、タグも生成されない。

#### 【0047】

特定の遺伝子を対象として、本発明の方法によって取得された遺伝子タグには、たとえば次のような有用性が期待できる。まず得られた遺伝子タグの塩基配列情報に基づいて、その遺伝子の転写開始点を明らかにすることができます。転写開始点は、全長cDNAの取得、あるいはプロモーターの探索において重要な情報である。たとえば5'側の塩基配列が明らかでないcDNAについて、本発明の方法を利用して、5'側のcDNAを取得することができる。あるいはすでに翻訳開始点が同定されている遺伝子であっても、その5'側の非翻訳領域(5'UTR)が完全なのかどうかを、遺伝子タグの情報によって評価することができる。

#### 【0048】

更に、同一のアミノ酸配列をコードしながら、転写開始点の異なる複数の転写産物を与える遺伝子が明らかにされている。ある遺伝子を対象に、さまざまなmRNAソースについて、本発明の遺伝子タグを取得すれば、当該遺伝子のあらゆる転写産物の転写開始点の情報を容易に集めることができる。もしも複数種類の遺伝子タグが得られれば、当該遺伝子には、転写開始点の異なる複数の転写産物が存在している可能性がある。すなわち本発明は、次の工程を含む転写開始点の異なる複数の転写産物の検出方法を提供する。

- (1) 本発明に基づいて遺伝子タグを取得する工程であって、cDNAの第1鎖の合成用のプライマーとして、解析すべき遺伝子に特異的なプライマーを用いる工程、
- (2) (1)で得られた遺伝子タグの塩基配列を比較する工程、および
- (3) 複数種の遺伝子タグが検出されたときに転写開始点の異なる複数の転写産物が検出される工程

本発明において検出された複数種の遺伝子タグと、前記遺伝子特異的プライマーの情報を利用して、各転写産物の転写開始点の塩基配列を決定することができる。更に、本発明に基づいて、各転写産物の発現レベルを比較することもできる。すなわち本発明は、次の工程を含む転写開始点の異なる複数の転写産物の発現レベルを比較する方法を提供する。

- (1) 本発明に基づいて遺伝子タグを取得する工程であって、cDNAの第1鎖の合成用のプライマーとして、解析すべき遺伝子に特異的なプライマーを用いる工程、
- (2) (1)で得られた遺伝子タグの塩基配列を比較する工程、および
- (3) 各遺伝子タグの出現頻度に基づいて、転写開始点の異なる複数の転写産物の発現レベルとして取得する工程

#### 【0049】

このほか、共通の塩基配列を有するRNAを意図的にcDNAとして合成することもできる。たとえば、保存性の高い蛋白質の機能ドメインを構成するアミノ酸配列に対して、それを

コードすると予測される塩基配列をもとに第1鎖合成用のプライマーをデザインすることができる。このプライマーを用いて合成されるcDNAは、特定の機能ドメインをコードする遺伝子のcDNAである可能性が高い。その結果、特定の機能ドメインを含む遺伝子のタグを意図的に集めることができる。こうして取得された遺伝子タグの発現レベルを比較することによって、特定の機能を有する遺伝子群の発現レベルを比較することができる。

#### 【0050】

いざれにせよ、本発明において合成されるcDNAの第1鎖は、その3'末端に、RNAリンカーに相補的な塩基配列を有している。したがって、この領域にアニールすることができるオリゴヌクレオチドを利用すれば、容易にcDNAの第2鎖を合成することができる。第2鎖の合成に先立って、第1鎖の鑄型としたRNAは、アルカリ加水分解によって除去することができる。本発明においては、第2鎖は、少なくとも、RNAリンカーに含まれるII<sub>s</sub>型の制限酵素の認識配列を含むように合成されるべきである。そのためには、たとえば、RNAリンカーの3'末端に配置されたII<sub>s</sub>型制限酵素の認識配列に相当する領域よりも、3'側において相補鎖合成を開始することができるプライマーを利用することができる。あるいは、II<sub>s</sub>型制限酵素の認識配列を含むプライマーを利用することもできる。

DNAを鑄型としてプライマー伸長反応によって相補鎖を合成する方法は公知である。すなわち、鑄型依存性のDNAポリメラーゼを利用して、相補鎖を合成する方法が知られている。DNAポリメラーゼとしては、T4 DNAポリメラーゼ、あるいはTaqポリメラーゼなどを用いることができる。

#### 【0051】

cDNAの合成に用いるプライマーは、任意の塩基配列を含むことができる。たとえばその5'末端側に、制限酵素の認識配列を付加したプライマーを利用することができる。プライマーの5'末端に、クローニングサイトを付与するための塩基配列を付加することは、広く行われている。

#### 【0052】

本発明において、cDNAの第2鎖は、固相に結合することができる標識を有するか、または固相に固定化されたプライマーによって合成することができる。プライマーを固相に結合することで、cDNAの第2鎖を固相に捕捉することができる。固相に捕捉されたcDNAは容易に回収することができる。

#### 【0053】

プライマーとして用いるオリゴヌクレオチドを固相に結合するための方法は任意である。たとえば、クロスリンカーを使ってオリゴヌクレオチドの5'末端をプレートに共有結合させる方法等が公知である（米国特許5656462）。あるいは、オリゴヌクレオチドを構成する塩基にビオチンのような結合親和性を持つ分子を導入することができる。ビオチンを、固相化したアビジンに結合させることによって、オリゴヌクレオチドは間接的に固相に捕捉される。オリゴヌクレオチドにおける、結合親和性分子の導入位置は制限されない。

#### 【0054】

第2鎖の合成によって2本鎖となったcDNAはII<sub>s</sub>型制限酵素で処理され、本発明における遺伝子タグが生成される。この段階で遺伝子タグは、RNAリンカーとして付加した塩基配列に連結された状態で回収することができる。遺伝子タグの回収のために、第2鎖合成用のプライマーが結合する固相が利用される。すなわち、遺伝子タグを結合した固相として回収される。固相は、II<sub>s</sub>型制限酵素を作用させる後、あるいは前に回収することができる。

#### 【0055】

さて、本発明における遺伝子タグの塩基配列を決定することによって、RNAの5'末端の塩基配列情報を得ることができる。遺伝子タグの塩基配列を決定する方法は任意である。しかし大量の遺伝子タグの塩基配列を効率的に決定するためには、SAGEの原理が有用である。すなわち、複数の遺伝子タグを連結させてコンカテマーとし、コンカテマーをクローニングして、複数のタグの塩基配列を一挙に決定することができる。

各遺伝子タグの長さは、タグの生成に用いたII<sub>s</sub>型制限酵素の作用によって一定である

と見なされる。したがって、コンカテマーは、一定の長さの遺伝子タグの塩基配列の繰り返しで構成されていると考えられる。そのため、コンカテマーの塩基配列から、各タグの塩基配列情報を得ることができる。

[0 0 5 6]

タグを連結してコンカテマーを得るための方法として、いくつかのバリエーションを示すことができる。以下のその例を述べる。まず広く知られているSAGEの原理を応用した方法について説明する。この方法においては、まず2つの遺伝子タグを向かい合わせに連結させてダイタグ(di-tag)を得る。このとき、もしもII型制限酵素による切断部分が粘着末端(sticky end)であるときは、予め平滑化しておく。平滑末端を形成するためには、T4 DNAポリメラーゼを作用させればよい。

次に、複数のダイタグを連結してコンカテマーを生成する。ダイタグを得るために、同じcDNAライブラリーを2つのプールにわけて、それぞれのプールに対して同じ操作で遺伝子タグを生成する。次に、2つのプール由来の遺伝子タグ同士を互いに連結してダイタグとする。このとき、遺伝子タグは、II型制限酵素で切断された切断部分で連結される。遺伝子タグは、T4DNAライガーゼなどによって酵素的に連結することができる。

【0 0 5 7】

ここで得られるダイタグは、以下の構造を有する。

PCR →

(固相) - [RNAリンカー] - [タグ] - [タグ] - [RNAリンカー] - (固相)  
 ←PCR

この段階で、ダイタグは、PCRなどの增幅方法によって増幅することができる。2つのプールの間でRNAリンカーの塩基配列が相違するようにしておけば、プールの異なるタグ間で連結されたダイタグが特異的に増幅されるので、タグ間の数的なバランスの崩れを防ぐことができる。本発明において、ダイタグの増幅は任意である。

【0058】

続いて複数のダイタグを連結してコンカテマーを得る。そのためには、たとえば予めRNAリシンカー内に制限酵素の認識配列を配置してくことができる。ダイタグを制限酵素で消化後に、制限酵素の切断部位をライゲーションすれば、複数のダイタグを連結することができる。こうして得られるコンカテマーの構造は、次のように示すことができる。

...../[Tag][Tag]/[Tag][Tag]/[Tag][Tag]/[Tag][Tag]/[Tag][Tag]/.....

すなわち、2つのタグを連結したダイタグ”[Tag][Tag]”を1単位として、制限酵素（アンカーリング酵素）による切断部位”/”を挟んでダイタグが連続した構造である。

【0 0 5 9】

更に、クローニング用ベクターの同じ制限酵素サイトに、コンカテマーをインサートすることができる。こうしてコンカテマーをインサートとして有するクローニングベクターを得ることができる。クローニングベクターのインサートの塩基配列を決定することによって、その中に含まれるタグの塩基配列が明らかになる。なおコンカテマーの長さは、1度のシーケンス反応で塩基配列を決定できる程度の長さであることが好ましい。たとえば、500 bp以下、たとえば20～400 bp、通常50～300 bpの範囲のコンカテマーを例示することができる。

[0 0 6 0]

ダイタグではなく、タグ単位で連結したコンカテマーを得ることもできる。たとえば、II<sub>s</sub>型制限酵素を作用させた後に、その切断部位にアダプターを結合することができる。このとき、タグは以下のような構造を有する。

PCR →

アダプターに制限酵素認識配列を配置しておけば、ダイタグのRNAリンカーを消化するのと同様にして、タグの両端を制限酵素で切断することができる。もしもタグを増幅する場合には、RNAリンカーとアダプターの塩基配列を利用してPCRによって増幅することもで

きる。いずれにしても、制限酵素で処理したタグを連結してコンカテマーとすることができる。コンカテマーは、更にクローニングベクターに組み込み、その塩基配列を明らかにすることができる。

II<sub>s</sub>型制限酵素によって切り出されるタグの長さは、ほぼ一定とされている。しかし、万が一、その長さにはばらつきがあると、ダイタグを構成したときに、正しいタグの塩基配列を同定することができない場合がある。ダイタグを経由しないでコンカテマーを構成すれば、万が一タグの長さが不均一であっても、タグの塩基配列を正確に決定することができる。

#### 【0061】

本発明の遺伝子タグの取得方法、更に取得されたタグの塩基配列の決定方法に必要な各種の試薬類は、予め組み合わせてキットとして供給することができる。すなわち本発明は、以下の要素を含む、遺伝子タグの製造用試薬キットに関する。

- (a) II<sub>s</sub>型制限酵素の認識配列を含むオリゴヌクレオチドからなるRNAリンカー
- (b) RNAリンカーをRNAのCAP部位に連結するための試薬
- (c) RNAリンカーを錫型として合成されたcDNAにアニールするオリゴヌクレオチドからなるcDNA第2鎖合成用のプライマー
- (d) cDNA第1鎖合成用プライマー

本発明のキットは、ダイタグやコンカテマーの調製に必要な試薬類を付加的に含むことができる。なお、これらの構成要素の具体的な構成は既に述べたとおりである。

#### 【0062】

本発明のキットにおいて、(d)cDNA第1鎖合成用プライマーとしては、たとえば以下のi)-iii)のいずれかに記載のプライマーを利用することができる。

- i) ランダムプライマー、
- ii) オリゴdTプライマー、および
- iii) 特定のmRNAに相補的な塩基配列を含むプライマー

試料に含まれる全てのmRNAを対象として遺伝子タグを製造する場合には、ランダムプライマー、あるいはオリゴdTプライマーが利用される。特にランダムプライマーは本発明における望ましいプライマーである。ランダムプライマーとは数十塩基の長さを有する不特定の塩基配列からなるオリゴヌクレオチドの集合体である。たとえば5～20、通常8～15塩基程度の長さのオリゴヌクレオチドが利用される。4種類の塩基の混合物を必要な長さに順次連結することにより、合成される。ランダムプライマーは、理論的には、あらゆる塩基配列に対して相補的な塩基配列を含んでいると考えることができる。

#### 【0063】

あるいは、特定のmRNAに相補的な塩基配列を含むプライマーによって本発明のキットを構成することもできる。特定のmRNAに特異的なプライマーを利用することによって、ある遺伝子の5'タグを特異的に製造することができる。こうして得られるタグの塩基配列情報を比較して、もしもその塩基配列にバリエーションが検出された場合には、当該遺伝子の転写産物には、5'末端の長さの異なる複数のバリエントが存在することが明らかにされる。したがって、特定のmRNAに相補的な塩基配列含むプライマーによって構成される本発明のキットは、特定の遺伝子の転写産物のバリエントを検出するためのキットとして有用である。

#### 【0064】

たとえば、以下のような要素によって、本発明の方法を実施するためのキットを構成することができる。各要素には、それぞれの要素を用いた反応に好適な緩衝液を添付することもできる。更に、本発明のキットには、遺伝子タグの塩基配列の解析のためのソフトウェアを組み合わせることもできる。

RNAリンカーを連結するための要素：

- ・BAP
- ・TAP
- ・T4 RNAリガーゼ

- ・RNAリンカー

cDNAの合成と分離のための要素：

- ・逆転写酵素
- ・DNAポリメラーゼ
- ・dXTP
- ・cDNA第1鎖合成用ランダムプライマー
- ・cDNA第2鎖合成用5'ビオチン化cDNA合成用プライマー
- ・アビジン結合磁気ビーズ

遺伝子タグを生成するための要素：

- ・II型制限酵素

ダイタグの生成と解析のための要素：

- ・T4 DNA リガーゼ
- ・遺伝子タグ増幅用プライマー
- ・DNAポリメラーゼ
- ・II型制限酵素
- ・シーケンシング用ベクター
- ・ベクターを形質転換するための宿主
- ・宿主を培養するための培地

### 【0065】

本発明によって生成されるコンカテマーの塩基配列情報の解析には、コンピューターソフトウェアを利用するのが有利である。たとえば以下のステップを実行することができるソフトウェアを、コンカテマーの塩基配列情報の解析に利用することができる。

シーケンサーの解析データを読み込むステップ

読み込まれた塩基配列データのタグ以外の塩基配列情報を識別するステップ

タグの塩基配列情報を蓄積するステップ

ここで、タグ以外の塩基配列情報としては、タグの形成過程で連結されたRNAリンカーやアダプターなどの塩基配列情報を示すことができる。あるいは、クローニングベクターに由来する塩基配列が読み取られる場合もあるかもしれない。いずれにせよ、これらの塩基配列情報は予め明らかな情報である。更に、これらの付加的な塩基配列情報とタグの塩基配列情報は規則的にコンカテマー上に配置されている。したがって、これらの塩基配列とタグの塩基配列とは、機械的に識別することができる。

### 【0066】

次にタグの塩基配列と認識された塩基配列情報が蓄積される。ダイタグを形成した場合には、アンチセンス鎖の塩基配列が読み取られる場合もあるので、相補配列の情報も合わせて記録する。アダプターを使ってダイタグを経由しないでコンカテマーを作成する場合には、アダプターとRNAリンカーのクローニングサイトを異なる配列となるようにデザインすれば、單一方向にクローニングすることができる。この場合には、相補配列の蓄積が必要でない。

### 【0067】

このプログラムには更に付加的な機能を持たせることができる。たとえば、得られたタグの塩基配列を比較し、同じ塩基配列を1つにまとめて、その出現頻度を記録するステップを実行させることができる。更に、異なるRNAソースのタグ情報を比較して、出現頻度の異なるタグを抽出するステップを実行させることもできる。

タグ情報の比較対象としては、予め集積されたデータベースの情報を利用することができる。たとえば、標準的な組織や細胞株について、予め本発明の方法に基づいて遺伝子タグの情報を集積しておく。この情報を、コンピューターネットワーク上で共有することができる。あるいは、前記試薬キットに添付して商業的に流通させることもできる。こうして入手された遺伝子タグ情報と、自身が実験して取得した遺伝子タグ情報を比較することもできる。

### 【0068】

本発明によって、転写産物であるmRNAの5'末端の塩基配列情報を得ることができる。5'末端の塩基配列情報は、遺伝子解析において、特に重要な意味を有する。たとえば、本発明によって得ることができる5'末端の塩基配列情報を、以下のような用途に利用することができる。

#### 【0069】

まず本発明は、遺伝子の発現プロファイルの取得に利用することができる。すなわち本発明は、次の工程を含む、真核細胞における遺伝子の発現プロファイルの取得方法に関する。

- (1)本発明に基づいて遺伝子タグを製造する工程、
- (2)(1)の遺伝子タグの塩基配列を決定する工程、および
- (3)決定された塩基配列とその出現頻度を対応付けることによって発現プロファイルを得る工程

#### 【0070】

本発明において、(1)遺伝子タグを製造する工程は、以下の工程を含むことができる。特に断らない場合には、以降の記載においても同様に、「本発明に基づいて遺伝子タグを製造する工程」とは、以下の工程を含む。

- (A) RNAのCAP部位にII<sub>s</sub>型制限酵素の認識配列を含むRNAリンカーを連結する工程、
- (B)(A)のRNAを錆型としてcDNAを合成する工程、
- (C)(B)のcDNAにRNAリンカーに含まれる認識配列を認識するII<sub>s</sub>型制限酵素を作用させ、遺伝子タグを生成する工程

#### 【0071】

一般に、発現プロファイルとは、発現情報を伴った遺伝子情報のリストを指す。発現情報とは、発現のレベルを示す量的なパラメーターである。遺伝子情報とは、通常、遺伝子を特定するための情報を言う。具体的には、遺伝子の塩基配列、遺伝子の名称、遺伝子のID番号などが遺伝子情報を構成する。リストを構成する遺伝子の数は、任意である。またその対象も、限定されない。解析の目的に応じて、必要な遺伝子の情報を集積して発現プロファイルが構成される。

#### 【0072】

本発明によれば、CAP構造を有するRNAから、その5'末端の塩基配列情報をタグ情報として取得することができる。またその塩基配列情報を照合し、同じ塩基配列の数をカウントすることによって、塩基配列情報とその出現頻度とが対応付けられる。こうして発現プロファイルを得ることができる。

#### 【0073】

RNAとして全てのRNAを対象とすれば、全遺伝子を対象とする発現プロファイルを得ることができる。本発明においては、特定の遺伝子、あるいは構造的な共通性を有する一群の遺伝子を対象に、遺伝子タグを生成することもできる。このようなケースでは、特定の遺伝子、あるいは一群の遺伝子の発現プロファイルが生成される。

#### 【0074】

CAP構造を有するmRNAとは、細胞中で発現しているmRNAの全てであると仮定すると、本発明によって得ることができる発現プロファイルは、細胞内の遺伝子の発現状況をより正確に反映していると言うことができる。本発明において、塩基配列の出現頻度をカウントするとき、解析対象となる塩基配列情報の総数に占めるある配列の出現頻度の相対的な数を蓄積するのが好ましい。特にPCRなどで増幅された後の出現頻度情報は、定量的な意味は小さい。総数に対する比として比較すれば、より客観的な評価を期待できる。

#### 【0075】

本発明によって得られた発現プロファイルは、データベースとすることができる。データベースとは、発現プロファイルを構成する情報を機械可読式のデータとして蓄積した電子データを言う。本発明のデータベースは、少なくとも、タグの塩基配列情報と、それに関連付けられた出現頻度情報を含む。更に本発明のデータベースは、各塩基配列情報のID番号、塩基配列情報が得られたRNAの由来を合わせて記録することができる。更に、既知

の遺伝子の塩基配列情報との関係、ゲノム上へのマッピングの結果などの情報を付加することもできる。

【0076】

本発明の発現プロファイルのデータベースは、電子媒体に保存することができる。電子媒体としては、各種のディスク装置、テープ媒体、あるいはフラッシュメモリーなどを示すことができる。これらの電子媒体は、ネットワーク上で共有することができる。たとえば、インターネット上で本発明のデータベースを共有することができる。更に、前記タグ配列の解析のためのソフトウェアに、インターネットを介して、本発明のデータベースの情報を参照するための機能を追加することもできる。あるいは逆に、本発明に基づいて生成された新たな発現プロファイル情報を、インターネットを介して、データベースに追加することもできる。

【0077】

本発明の発現プロファイルを利用して、発現プロファイル解析を実施することができる。すなわち本発明は、本発明に基づいて異なる種類の細胞の遺伝子発現プロファイルを取得し、遺伝子発現プロファイルを比較して細胞間で発現頻度の異なる遺伝子タグを選択する工程を含む、遺伝子発現プロファイルの解析方法に関する。異なる細胞間で発現レベルの異なる遺伝子を取得する解析方法は、発現プロファイル解析と呼ばれている。このような解析によって、たとえば、疾患などに関連する遺伝子が数多く取得されてきた。本発明の発現プロファイルも、このような発現プロファイル解析に利用することができる。

【0078】

本発明の発現プロファイル解析において、解析の対象とする異なる細胞とは、その由来が異なるあらゆる細胞を言う。同じ組織に由来する細胞であっても、疾患の有無、人種、年齢、性別などのなんらかの条件の相違がある場合には、由来が異なる細胞である。解析の目的の応じて、考慮すべき条件が相違すれば、由来が異なる細胞である。一方、解析の目的に対して無視しうる条件の相違しか見出せない場合には、同一の細胞と見なされる。たとえば、異なる臓器、異なる組織、あるいは由来や培養条件などが異なる細胞の間で発現プロファイルを比較することによって、臓器、組織、あるいは細胞間において、発現レベルの高い（または低い）遺伝子を選択することができる。本発明を応用することができる、解析対象の組み合わせを以下に例示する。

【0079】

異なる組織

成人の組織と胎児の組織

患者の組織と健常者の組織

男性の組織と女性の組織

人種の異なるヒトの組織

生育環境の異なる同じ生物種の組織

異なる細胞

同じ細胞で培養条件の異なる細胞

同じ培養条件で培養時間の異なる細胞

特定の処理を与えた細胞と与えない細胞

【0080】

より具体的には、癌組織と、正常な組織の間で発現プロファイルを比較することによって、癌に特徴的な遺伝子タグを取得することができる。あるいは、特に悪性度の高い癌と、低い癌との比較によって、悪性度に関連する遺伝子タグを特定することができる。

【0081】

本発明によって得られる遺伝子タグは、mRNAの5'末端の塩基配列情報を含んでいる。したがって、同じ蛋白質をコードする遺伝子であって、5'UTRの構造が異なるバリエントを、異なる転写産物として発現プロファイルに反映させることができる。この特徴は、公知のSAGEによって得ることができるタグと比較して、本発明のタグが有している大きなメリットの一つである。また本発明の遺伝子タグは、タグの塩基配列情報そのものが全長cDNA

の5'側のプライマーの塩基配列情報として有用である。したがって、発現プロファイル解析によってピックアップしたタグの塩基配列情報に基づいてデザインしたプライマーと、オリゴdTプライマーを利用すれば、直ちに全長cDNAを合成することができる。あるいは、mRNAの任意の領域に相補的な塩基配列を有するプライマーを組み合わせれば、mRNAの5'側の塩基配列を含むcDNAを取得することができる。このことも本発明の大きな特徴である。

#### 【0082】

本発明によって得ることができる遺伝子タグは転写産物であるmRNAの5'末端の塩基配列を含んでいる。したがって、この塩基配列をゲノムの塩基配列上にマッピングすることによって、遺伝子の転写開始点を同定することができる。すなわち本発明は、次の工程を含む、遺伝子の転写開始点の決定方法に関する。

- (1) 本発明の方法に基づいて遺伝子タグを製造する工程、
- (2) (1)の遺伝子タグの塩基配列を決定する工程、および
- (3) 決定された塩基配列をゲノムの塩基配列上にマッピングし、塩基配列が一致した領域を当該遺伝子の転写開始点として同定する工程

#### 【0083】

2003年4月、ヒトゲノムシークエンス国際コンソーシアムは、ヒトゲノムの解読完了を発表した。この結果、全ゲノムの 99% (28億3000万塩基対) を 99.99% の精度でカバーするヒトゲノム精密配列を手にすることができます。一方、本発明は細胞内で転写されているあらゆるmRNAの5'末端をタグとして生成する。したがって、原理的には、ある細胞において転写されている遺伝子の、ほぼ全ての転写開始点をゲノム上にマッピングすることができる。ゲノム上にマッピングされた転写開始点は、転写調節領域の取得において重要な情報である。

#### 【0084】

たとえば、転写開始点の上流の1～2 kbの範囲をクローニングし、転写調節因子のスクリーニングに利用することができる。あるいは、この領域の塩基配列を解析することによって、転写調節領域を予測することもできる。より具体的には、既知の転写因子の認識配列が保存されている領域を探索することによって、転写因子の結合領域の予測が可能である。

#### 【0085】

また転写開始点のマッピングは、遺伝子そのもののマッピングに他ならない。つまり、本発明におけるタグの塩基配列情報のマッピングの結果に基づいて、遺伝子のゲノム上における物理的な位置関係を把握することができる。現状では、遺伝子の転写開始点は、質の高い全長cDNAの塩基配列情報に頼らなければマッピングすることはできなかった。ところが本発明によって得ることができるタグ情報を利用すれば、容易に転写開始点をマッピングすることができる。このように、本発明によって得ができるタグ情報は、全長cDNAの成果に匹敵する価値を有していると言えることができる。

#### 【0086】

加えて本発明によって得ができる遺伝子タグの塩基配列情報は、cDNAの全長率の評価に利用することができる。ゲノムの塩基配列が明らかにされる一方で、細胞の働きを蛋白質レベルで明らかにするための様々な試みが続けられている。そのための手法の一つとして、全長cDNAの網羅的な解析がある。全長cDNAの網羅的な解析においては、ある細胞で発現している遺伝子の全長が網羅的に取得され、その構造が決定される。このときに、取得されたcDNAの全長性が高いことが重要な条件となる。

#### 【0087】

まず第1に、少なくともORFを特定するために、mRNAの5'側の塩基配列が明らかにされている必要がある。更に、転写開始点を同定するためには、5'末端まで取得されていることが重要である。これらの条件を満たしていることを確認するために、しばしば得られたcDNAの全長性が評価される。cDNAの全長性とは、mRNAの5'末端の塩基配列を含むcDNAが、取得されたcDNA全体のどの程度を占めているかを表すパラメーターである。

#### 【0088】

本発明の遺伝子タグは、mRNAの5'末端の塩基配列情報を提供する。したがって、網羅的に取得されたcDNAの塩基配列と、同じライブラリーから得られた本発明の遺伝子タグの塩基配列を照合することによって、各cDNAの5'末端がmRNAの5'末端の塩基配列を含むかどうかを明らかにすることができます。もしも遺伝子タグの塩基配列の多くが、cDNAの塩基配列上にマッピングできる場合には、取得されたcDNAの多くが全長である可能性が高い。逆に、遺伝子タグと一致する塩基配列が取得されたcDNA中に見出せない場合には、cDNAの全長性は低いと予測される。

### 【0089】

本発明における遺伝子タグの塩基配列情報は、mRNAの5'末端の塩基配列を含むcDNAの取得に利用することができる。すなわち本発明は、次の工程によって決定された塩基配列またはその相補配列を含むcDNAを合成するための5'側のプライマーと、cDNAの任意の部位にアニールする3'側のプライマーを含む、cDNA合成用プライマーセットに関する。

- (1)本発明に基づいて遺伝子タグを製造する工程、および
- (2)(1)の遺伝子タグの塩基配列を決定する工程

### 【0090】

本発明のプライマーセットを構成する5'側プライマーの塩基配列は、タグとして取得された塩基配列、またはその相補配列を含む。タグは、mRNAのセンス配列、あるいはアンチセンス配列として得られる。したがって、その相補配列あるいは、タグの塩基配列そのものが、cDNA合成用の5'側のプライマーの塩基配列として利用される。5'側のプライマーが5'末端において相補鎖合成を開始することから、本発明のプライマーセットによって合成されるcDNAが常に5'末端の塩基配列を含む。なおタグ配列はDNAから得られるので、塩基tを含む。これに対してRNAの5'末端配列は、tに相当する塩基がuであることは言うまでもない。

### 【0091】

一方本発明のプライマーセットを構成する3'側のプライマーには、cDNAにアニールすることができる任意のプライマーを利用することができる。3'側のプライマーの選択によって、様々なcDNAを合成することができる。本発明プライマーセットに利用することができる3'側のプライマーとしてたとえば次のようなプライマーを示すことができる。

- i)オリゴdTプライマー
- ii)cDNAの断片配列情報、および
- iii)cDNAのII型制限酵素認識に隣接する遺伝子タグの塩基配列またはその相補配列からなるプライマー

### 【0092】

まずオリゴdTプライマーとの組み合わせは、全長cDNAの合成に有用である。次に、cDNAの断片配列情報に基づいてデザインされた3'側プライマーは、当該cDNAの5'側の領域を取得するためのプライマーとして利用される。このような目的のためには、できるだけ当該cDNAの5'側の塩基配列に基づいて、3'側プライマーをデザインすればよい。cDNAの断片情報にはESTが含まれる。また様々な遺伝子解析によって、cDNAの断片情報が取得される。そして、しばしば断片情報に基づいて全長の塩基配列を決定することが試みられる。たとえば、DNAアレイのプローブとして使われているESTの5'側の塩基配列の取得が必要なとき、本発明のプライマーセットを利用して、目的とする領域を合成することができる。あるいは、PCRクローニングなどによって取得されたcDNAの断片から、その全長の取得を試みる場合もある。本発明において、cDNAの断片配列情報とは、特定のmRNAに相補的な塩基配列を含むプライマーと定義することができる。

### 【0093】

更に、cDNAのII型制限酵素認識に隣接する遺伝子タグの塩基配列またはその相補配列からなるプライマーを3'側プライマーとして利用することもできる。現在実用化されているSAGEは、cDNA中に含まれる特定の制限酵素サイトに隣接する領域を遺伝子タグとして生成する。このタグの塩基配列情報に基づいて、遺伝子発現プロファイルを解析することができる。同じ解析対象について、既知の解析方法に基づいて選択された遺伝子タグの塩基配

列情報を3'側のプライマーとして利用すれば、着目する遺伝子のかなりの部分を含むcDNAを合成できる可能性がある。

#### 【0094】

これらのプライマーセットのうち、オリゴdTプライマーとの組み合わせは、全長cDNAを合成するためのプライマーセットとして特に好ましい。全長cDNAは、転写開始点のマッピングに有用である。また5'UTRの構造が異なる転写産物の同定のためには、少なくとも5'末端を含む領域の塩基配列の決定が必須である。更に、全長cDNAは、通常は取得が難しいとされている。こうした背景から、本発明に基づいて得られた遺伝子タグ情報をを利用して、全長cDNAを合成することは、特に有用性が大きい。すなわち本発明は、次の工程を含む、全長cDNAの合成方法に関する。

a)次の工程によって決定された塩基配列またはその相補配列を含むcDNAを合成するための5'側のプライマーと、オリゴdTプライマーからなる3'側のプライマーを用い、RNAあるいはcDNAを鑄型として相補鎖合成反応を行う工程、および

(1)本発明の方法に基づいて遺伝子タグを製造する工程、および

(2)(1)の遺伝子タグの塩基配列を決定する工程

b)合成されたDNAを全長cDNAとして回収する工程

#### 【0095】

目的とするmRNAを含む可能性の高い細胞から取得されたRNAを鑄型として、前記の本発明のプライマーセットを用いてcDNAが合成される。あるいは当該細胞から得られたcDNAライブラリーを鑄型として利用することもできる。当業者は、与えられたプライマーの塩基配列情報に基づいて、cDNAを合成することができる。具体的には、RT-PCRなどの公知の方法を利用して、RNAから目的とするcDNAを合成することができる。RNAはmRNAを用いるのが好ましい。あるいはcDNAライブラリーを鑄型とするときは、PCRによって目的とするcDNAを合成することができる。cDNAライブラリーには市販のライブラリーを用いることもできる。

#### 【0096】

本発明は、こうして合成された全長cDNAに関する。本発明において、全長cDNAとは、mRNAのCAP構造を有する部分の塩基配列情報と、poly(A)を含むcDNAを言う。本発明はまた、本発明に基づいて合成された全長cDNAによってコードされるポリペプチドに関する。全長cDNAの塩基配列を解析し、ORFを同定することができる。同定されたORFに基づいて、コード領域を発現ベクターに導入することができる。本発明は、このようにして得ることができる発現ベクターを含む。当該発現ベクターを適当な発現系に導入して、cDNAによってコードされるポリペプチドを組み換え体として発現させ、更に回収することができる。

#### 【0097】

また本発明の全長cDNAのコード領域がコードするポリペプチドは、生体外翻訳(in vitro translation)によって組み換え体として発現させ、回収することができる。生体外翻訳の方法は公知である。生体外翻訳は無細胞タンパク質翻訳とも呼ばれている。すなわち、目的とするアミノ酸配列をコードするDNAをプロモーターに機能的に連結したコンストラクトを、生体外翻訳を支持する要素と接触させることにより、アミノ酸配列に翻訳することができる。コンストラクトには、ターミネーターなどの転写調節領域を配置することもできる。生体外翻訳を支持する要素は、RNAポリメラーゼ、リボヌクレオチド基質、アミノ酸、リボソーム、およびtRNAなどを含む混合物である。これらのタンパク質翻訳に必要な成分が存在すれば、細胞機能を利用することなく、DNAをタンパク質に翻訳することができる。RNAポリメラーゼは、前記プロモーターを認識してその制御下のDNAを鑄型としてmRNAに転写する。転写にはリボヌクレオチド基質ATP、GTP、CTP、およびUTPが用いられる。転写されたmRNAはリボソームにおいてポリペプチドに翻訳される。

生体外翻訳を支持する要素として、市販のin vitro translation用のキットを用いることができる。ウサギ網状赤血球のライセート(Rabbit Reticulocyte Lysate ; RRL)、小麦胚芽抽出物(Wheat Germ Extract ; WGE)、あるいは大腸菌のライセートなどを利用した無細胞タンパク質翻訳のためのキットが市販されている。あるいは転写、翻訳およびエネル

ギー再生に必要な約30の酵素類をそれぞれ高純度で精製後、再構成したin vitro 転写・翻訳システムも実現され (Shimizu et al. (2001) *Nature Biotechnology*. vol.19, p. 751-755)、キットとして商業的に供給されている。

#### 【0098】

加えて本発明は、当該ポリペプチドを認識する抗体に関する。抗体は、たとえば前記組み換え体、あるいは翻訳アミノ酸配列から選択されたアミノ酸配列からなるドメインペプチドで免疫動物を免疫することによって得ることができる。免疫動物からはポリクローナル抗体を回収することができる。更に、免疫動物の抗体産生細胞をクローニングして、モノクローナル抗体を得ることができる。抗体産生細胞をミエローマのような細胞株と融合させて、ハイブリドーマとし、目的とする反応性を有する抗体を産生するクローンをスクリーニングするための方法が公知である。

以下に、実施例に基づいて、本発明を更に具体的に説明する。

#### 【実施例1】

##### 【0099】

本発明に基づいて、mRNAの5'末端の塩基配列を含む遺伝子タグを取得できることを以下の実験によって確認した。以下の操作の概略を図1に示した。

##### オリゴキャップ法

オリゴキャップ法は、Maruyama および Sugano (1994)の方法を改変して行った (Maruyama, K., Sugano, S., 1994. Oligo-capping: a simple method to replace the cap structure of eucaryotic mRNAs with oligoribo-nucleotides. *Gene* 138, 171-174.)。5-10  $\mu$ gのポリ(A)+ RNAを、100ユニットのRNasin (Promega) を添加した総液量100  $\mu$ lの100 mM Tris-HCl (pH 8.0) および5 mM 2-メルカプトエタノール混合液中で、1.2ユニットのバクテリア由来アルカリフィオスファターゼ(BAP;TaKaRa)により37°C、40分間処理した。フェノール:クロロフォルム(1:1)抽出処理を2回行い、エタノール沈殿処理した。得られた該ポリ(A)+ RNAを100ユニットのRNasinを添加した総液量100  $\mu$ lの50 mM 酢酸ナトリウム(pH 5.5)、1 mM EDTA、5 mM 2-メルカプトエタノール混合液中で、20ユニットのタバコ酸性ピロホスファターゼ(TAP)により37°C、45分間処理した。

##### 【0100】

フェノール:クロロホルム抽出処理およびエタノール沈殿処理の後、2-4  $\mu$ gのBAP-TAP処理ポリ(A)+ RNAを2つのプールに分けて、各プールをRNAリンカー (5'-oligo 1および5'-oligo 2)とそれぞれライゲーションさせた。5'-oligo 1および5'-oligo 2は、それぞれ次の塩基配列を有するRNAである。いずれのRNAリンカーも、XhoIおよびMmeI認識配列を含む。

5'-oligo 1/配列番号：1

5'-UUU GGA UUU GCU GGU GCA GUA CAA CUA GGC UUA AUA CUC GAG UCC GAC -3'

5'-oligo 2/配列番号：2

5'-UUU CUG CUC GAA UUC AAG CUU CUA ACG AUG UAC GCU CGA GUC CGA C -3'

250ユニット RNA ligase (TaKaRa)、および100ユニット RNasinを、下記組成の反応混合液で総液量100  $\mu$ Lとし、20°C、3-16時間反応させ、RNAリンカーをライゲーションした。

50 mM Tris-HCl (pH 7.5)

5 mM MgCl<sub>2</sub>

5 mM 2-メルカプトエタノール

0.5 mM ATP

25% PEG8000

##### 【0101】

##### cDNA合成

cDNAの合成にあたり、完全長cDNA富化ライプラリーと5'末端cDNA富化ライプラリーの2種類のライプラリーを合成した。完全長cDNA富化ライプラリーは、オリゴdTアダプタープライマーを使ってpoly(A)+mRNAを錆型として合成されたcDNAからなる、完全長cDNAに富む

ライプラリーである。一方、5'末端cDNA富化ライプラリーは、cDNAの合成にランダムアダプタープライマーを使って合成されたcDNAからなっている。ランダムアダプタープライマーの使用によって、poly(A)を伴わない断片からも、cDNAが合成されている。これら2種類のcDNAのそれについて、遺伝子タグの取得を試みた。

#### 【0102】

ライゲーションされなかったRNAリンカーを取り除いた後、RNaseH フリーの逆転写酵素(Superscript II, Gibco BRL)によりcDNAを合成した。完全長cDNA富化ライプラリーを得るために、10 pmolのdTアダプタープライマー(配列番号：3)を、2-4 μgのオリゴキャップポリ(A)+ RNAを含む50 μlに加えてcDNAを合成した。

dTアダプタープライマー(配列番号：3)

5'-GCG GCT GAA GAC GGC CTA TGT GGC CTT TTT TTT TTT TTT-3'

反応条件はメーカー推奨の方法に従った(42°C、1時間インキュベート)。

#### 【0103】

更に5'末端cDNA富化ライプラリーを得るために、10 pmolのランダムアダプタープライマー(配列番号：4)を用いて12°C、1時間のインキュベーションを行い、更に42°C、1時間インキュベーションを行った。

ランダムアダプタープライマー(配列番号：4)

5'-GCG GCT GAA GAC GGC CTA TGT GGC CNN NNN NC-3'

#### 【0104】

cDNAの増幅

第1鎖cDNAを合成した後、RNAを15 mM NaOHで65°C、1時間処理することにより分解した。1 μgのオリゴキャップポリ(A)+ RNAを錆型として合成されたcDNAを、100 μl中に16 pmolの5'PCRプライマーおよび3'PCRプライマー(5'-GCG GCT GAA GAC GGC CTA TGT-3' / 配列番号：7)を含むXL PCRキット(Perkin-Elmer)を用いて増幅した。5'PCRプライマーは、RNAリンカーとして5'oligo-1をライゲーションしたプールについては配列番号：5の、また5'oligo-2のプールには配列番号：6のプライマーをそれぞれ用いた。

5'oligo 1用5'PCRプライマー/配列番号：5

5'ビオチン- GGA TTT GCT GGT GCA GTA CAA CTA GGC TTA ATA-3'

5'oligo 2用5'PCRプライマー/配列番号：6

5'ビオチン- CTG CTC GAA TTC AAG CTT CTA ACG ATG TAC G-3'

3'PCRプライマー(配列番号：7)

5'-GCG GCT GAA GAC GGC CTA TGT-3'

第1鎖の合成にdT-アダプタープライマーをプライマーとして用いた場合、94°C 1分間、58°C 1分間および72°C 10分間のサイクルを5~10回繰り返してcDNAの増幅を行った。また第1鎖の合成にランダムアダプタープライマーをプライマーとして用いた場合には、94°C 1分間、58°C 1分間および72°C 2分間のサイクルを10回繰り返してcDNAの増幅を行った。

#### 【0105】

PCR産物は、1回のフェノール:クロロフォルム(1:1)処理の後、エタノール沈澱処理を経て、MmeI型IIIs 制限酵素(University of Gdansk Center for Technology Transfer, Gdansk, Poland)により処理した。制限酵素処理は、総液量300 μlの10mM HEPES、pH8.0、2.5mM 酢酸カリウム、5mM 酢酸マグネシウム、2 mM DTTおよび40 μM S-アデノシルメチオニン混合液中で40ユニットのMmeIを用いて37°C、2.5時間行った。制限酵素処理された5'末端cDNA断片はストレプトアビジンでコートされたマグネットイックビーズ(Dynal, Oslo, Norway)に結合させた。4ユニットのT4 DNAリガーゼを添加した供給バッファーを含む16 μlの反応溶液中で16°C、2.5時間反応させて、ビーズに結合しているcDNA断片を互いに直接結合させてダイタグを得た。

#### 【0106】

生成したダイタグはプライマー5' -GGA TTT GCT GGT GCA GTA CAA CTA GGC- 3' (配列番号：8) および5' -CTG CTC GAA TTC AAG CTT CTA ACG ATG-3' (配列番号：9) を用

いて、PCRにより増幅した。PCR産物をポリアクリルアミドゲル電気泳動(PAGE)により確認し、XhoIにより処理した。ダイタグを含むバンドを切り出し、セルフライゲーションさせて長いコンカテマーを形成させた。このコンカテマーをpZero 1.0 (Invitrogen)のXhoI部位に組み込んだ。

【0107】

M13 フォワードプライマーおよびM13リバースプライマーを使用したPCRによりコロニーのスクリーニングを行った。600 bp以上のインサートを含むPCR産物は、Big Dye terminator ver.3を用いて、3730 ABI自動DNAシークエンサー(Applied Biosystems, CA)により配列を決定した。全ての電気泳動図に対して、不明瞭な塩基の有無を確認するため、およびミスリーディングを修正するために目視による再解析を行った。

【0108】

各タグの出現頻度を、そのために作製したソフトウェアで測定した。解析の結果得られたタグの塩基配列をqueryとして、BLASTサーチ (<http://www.ncbi.nlm.nih.gov/BLAST/>) およびthe human genome database (<http://www.ncbi.nlm.nih.gov/genome/guide/human/>)のデータを検索した。

【0109】

ランダムアダプタープライマーによって合成された5'末端cDNA富化ライブラリーから得られた3000以上の遺伝子タグの塩基配列を解析した結果の一部を以下にまとめた。以下の結果においては、遺伝子タグの塩基配列を記載した配列番号に続けて、次の情報を/で区切って記載した。またこれらの情報の後に行を改めてヒットした既知遺伝子の情報 (GenBank Accession No.とアノテーション) を記載した。

遺伝子タグの塩基配列

得られた遺伝子タグの総数における当該遺伝子タグの出現頻度

遺伝子タグの塩基配列がヒットした既知配列の位置 (○: 5'末端にヒットしたと考えられるもの、×: 5'末端の塩基配列でないと考えられたもの)

(配列番号: 1 0) / ACATCTGACCTCATGGAG / 27 / ○

gi|33694637|tpg|BK000408.1| TPA: Human adenovirus type 5, complete genome

(配列番号: 1 1) / CTCTTCCTGCCTAACG / 22 / ○

gi|17981705|ref|NM\_001007.2| Homo sapiens ribosomal protein S4, X-linked (RPS4X), mRNA

(配列番号: 1 2) / TACCTGGTTGATCCTGCC / 21 / ×

(配列番号: 1 3) / CTTTCCTGTGGCAGCAG / 20 / ○

<gi|16579884|ref|NM\_000968.2| Homo sapiens ribosomal protein L4 (RPL4), mRNA

(配列番号: 1 4) / CTCTCCGCCGTCGTCCG / 16 / ○

Homo sapiens eukaryotic translation elongation factor 2 (EEF2), mRNAの上流

(配列番号: 1 5) / CTCATTGAACTCGCCTGC / 11 / ○

gi|28338|emb|X04098.1|HSACTCGR Homo sapiens mRNA for cytoskeletal gamma-actin (ACTG1 gene)

(配列番号: 1 6) / CTGGTTGATCCTGCCAGT / 11 / ×

(配列番号: 1 7) / CTCAGTCGCCGCTGCCAG / 10 / ○

gi|28338|emb|X04098.1|HSACTCGR Homo sapiens mRNA for cytoskeletal gamma-actin (ACTG1 gene)

(配列番号: 1 8) / CTTTCACTGCAAGGCCGGC / 10 / ○

gi|18314626|gb|BC021993.1| guanine nucleotide binding protein (G protein), beta polypeptide 2-like 1

(配列番号: 1 9) / ACGCTGTGACAGCCACAC / 9 / ○

NM\_005382の上流

(配列番号: 2 0) / GTGACAGCCACACGCC / 9 / ×

gi|35045|emb|Y00067.1|HSNFM Human gene for neurofilament subunit M (NF-M)  
(配列番号：21) / AACGGCTAGCCTGAGGAG / 8 / ×

gi|188487|gb|M59828.1|HUMMHSP Human MHC class III HSP70-1 gene (HLA), complete cds  
(配列番号：22) / AGTAGCAGCAGCGCCGGG / 8 / ○

gi|14043071|ref|NM\_031243.1| Homo sapiens heterogeneous nuclear ribonucleoprotein A2/B1  
(配列番号：23) / ATTCCCTAGTTAAGGCGGC / 8 / ○

gi|5020073|gb|AF146651.1|AF146651 Homo sapiens glyoxalase-I gene, complete cds  
(配列番号：24) / AATTGTGTTCGCAGCCGC / 7 / ○

gi|22027640|ref|NM\_002107.2| Homo sapiens H3 histone, family 3A (H3F3A), mRNA  
(配列番号：25) / ATATTCTTACTCTCTCG / 7 / ×

gi|37704377|ref|NR\_001564.1| Homo sapiens X (inactive)-specific transcript (XIST) on chromosome X  
(配列番号：26) / CTCAGTCGCCGCTGCCAA / 7 / ○

gi|28338|emb|X04098.1|HSACTCGR Homo sapiens mRNA for cytoskeletal gamma-actin  
(配列番号：27) / AAAACGGCCAGCCTGAGG / 6 / ×

gi|188489|gb|M59830.1|HUMMHSP2 Human MHC class III HSP70-2 gene (HLA), complete cds  
(配列番号：28) / CTCTCTTCACTGCAAGG / 6 / ○

gi|12652914|gb|BC000214.1| guanine nucleotide binding protein (G protein), beta polypeptide 2-like 1  
(配列番号：29) / AATTCTACCGCGCACCGG / 5 / ○

gi|402305|gb|L24369.1|HMRPS4A Homo sapiens ribosomal protein S4 gene  
(配列番号：30) / ACCGCGAGACCGCGTCC / 5 / ○

gi|10437878|dbj|AK025375.1| Homo sapiens ACTB mRNA for mutant beta-actin  
(配列番号：31) / AGACGCAGACTAGATTGT / 5 / ○

gi|2315183|emb|Z82216.1|HS75N13 Human DNA sequence from clone RP1-75N13 on chromosome Xq21.1,  
(配列番号：32) / AGTCGATCGGTAGCGGG / 5 / ×

gi|37540535|ref|XM\_294582.2| Homo sapiens similar to DNA-binding protein B (L0 C347295), mRNA  
(配列番号：33) / AGTTCTCGGCGTACGGC / 5 / ○

gi|30581134|ref|NM\_006306.2| Homo sapiens SMC1 structural maintenance of chromosomes 1-like 1  
(配列番号：34) / AGTTGCTTCAGCGTCCCC / 5 / ○

gi|32487|emb|X15183.1|HSHSP90R Human mRNA for 90-kDa heat-shock protein  
(配列番号：35) / ATTAAACGGTTGCAGGCG / 5 / ×

gi|33239450|ref|NM\_182649.1| Homo sapiens proliferating cell nuclear antigen (PCNA) transcript variant 2, mRNA  
(配列番号：36) / CCGGCCGGGGGGCGGGCG / 5 / ○

gi|555853|gb|U13369.1|HSU13369 Human ribosomal DNA complete repeating unit  
(配列番号：37) / CCTTTGGCTCTCTGACC / 5 / ○

gi|15718688|ref|NM\_001006.2| Homo sapiens ribosomal protein S3A (RPS3A), mRNA  
(配列番号：38) / CTCAGTACAGCTCCGGCC / 5 / ○

gi|21217408|gb|AC015849.5| Homo sapiens chromosome 17, clone RP11-362K1, complete sequence  
(配列番号：39) / CTCTTCGGCCGCGCTGG / 5 / ○

gi|461248|dbj|D28421.1|HMRPL80 Homo sapiens mRNA for ribosomal protein L8 homologue, 5'UTR

## 【0110】

得られたタグのうち30の塩基配列の解析の結果、73%以上(22/30)のタグは、実際にcDNAの5'末端の塩基配列であった。本発明に基づいて、高い確率でmRNAの5'末端の塩基配列をタグとして取得できることが裏付けられた。

## 【実施例2】

## 【0111】

本発明に基づく、mRNAの5'末端の塩基配列を含む遺伝子タグを利用した遺伝子発現解析(以下、5'SAGEと記載する)の結果を公知のSAGE法(以下3'SAGEと記載する)と比較した。

## 材料および方法

## 3'-Long SAGEライブラリーの作製

HEK293から全RNAを単離し、前述のようにmRNAを選択した(Hashimoto, S.-i., Suzuki, T., Dong, H.-Y., Yamazaki, N. & Matsushima, K. Serial analysis of gene expression in human monocytes and macrophages. *Blood* 94, 837-844, 1999)。標準のSAGE手順を以下のように変更して使用し、mRNA 3μgでLong SAGE法(Saha, S. et al. Using the transcriptome to annotate the genome. *Nat Biotechnol* 20, 508-512, 2002)を行った。

## 【0112】

すなわち、NlaIII切断後に、リンカー1A (5'-TTT GGA TTT GCT GGT GCA GTA CAA CTA G GC TTA ATA TCC GAC ATG-3'／配列番号：40) とリンカー1B (5'-TCG GAT ATT AAG CCT AGT TGT ACT GCA CCA GCA AAT CC C7アミノ修飾-3'／配列番号：41) を互いにアニーリングし全cDNAの半分に連結し、リンカー2A (5'-TTT CTG CTC GAA TTC AAG CTT CTA ACG ATG TAC GTC CGA CAT G-3'／配列番号：42) とリンカー2B (5'-TCG GAC GCA CAT CGT TAG AAG CTT GAA TTC GAG CAG C7アミノ修飾-3'／配列番号：43) を互いにアニーリングしcDNAの残り半分に連結し、MmeI認識部位を含むリンカーを3'cDNA末端に連結した。MmeIタイプIIIS制限酵素(グダニスク大学技術移転センター(University of Gdansk Center for Technology Transfer)、ポーランド、グダニスク)を用いて、cDNAからリンカータグ分子を遊離させた。切断は、300μlの10 mM HEPES、pH 8.0、2.5 mM酢酸カリウム、5 mM酢酸マグネシウム、2 mM DTT、および40μM S-アデノシルメチオニン中でMmeI 40ユニットを使用し、37℃で2.5時間行った。供給緩衝液中にT4 DNAリガーゼ4ユニットを含む反応液16μl中で、リンカー1タグ分子とリンカータグ2分子を16℃で2.5時間、互いに直接連結させた。

## 【0113】

遊離したタグを互いに連結し、連鎖させ、pZero 1.0(インビトロジェン(Invitrogen))のSphI部位にクローニングした。M13フォワードおよびM13リバースプライマーを用いて、ポリメラーゼ連鎖反応法(PCR)によりコロニーをスクリーニングした。600 bpを超える挿入断片を含むPCR産物を、Big Dyeターミネーターver.2でシークエンシングし、3730 ABI自動DNAシークエンサー(アプライドバイオシステムズ(Applied Biosystems)、カリフォルニア州)を用いて解析した。電気泳動図はすべて目視検査により再度解析し、あいまいな塩基を調べ誤読を訂正した。SAGE 2000ソフトウェア(バージョン4.12)を用いて、各タグの存在量を定量した。リンカー配列、他の可能性のある人工産物、および反復したダイタグを除去した後、各タグを解析した。

## 【0114】

## 5'-SAGEライブラリーの作製

いくつかの変更点(Suzuki, Y., Yoshitomo-Nakagawa, K., Maruyama, K., Suyama, A. & Sugano, S. Construction and characterization of a full length-enriched and a 5'-end-enriched cDNA library. *Gene* 200, 149-156, 1997)を加え、MaruyamaおよびSugano(Maruyama, K. & Sugano, S. Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. *Gene* 138, 171-174, 1994)に記載されるように、オリゴキャップ法を行った。

## 【0115】

つまり、RNasin (プロメガ (Promega) ) 100ユニットを含む100 $\mu$ lの100 mM Tris-HCl (pH 8.0) 、5 mM 2-メルカプトエタノール中で、ポリ(A)+ RNA 5~10 $\mu$ gを細菌由来アルカリホスファターゼ (BAP; タカラ (TaKaRa) ) 1.2ユニットを用いて37°Cで40分間処理した。フェノール：クロロホルム (1:1) で2回抽出しエタノール沈殿した後、RNasin 100ユニットを含む100 $\mu$ lの50 mM酢酸ナトリウム (pH 5.5) 、1 mM EDTA、5 mM 2-メルカプトエタノール中で、ポリ(A)+ RNAをタバコ酸性ピロホスファターゼ (TAP) 20ユニットを用いて37°Cで45分間処理した。フェノール：クロロホルム抽出しエタノール沈殿した後、BAP-TAP処理したポリ(A)+ RNA 2~4 $\mu$ gを2つのプールに分け、XhoI / MmeI認識部位を含む以下のRNAリンカーの1つを各プールに連結した：RNasin 100ユニットを含む100 mlの50 mM Tris-HCl (pH 7.5) 、5 mM MgCl<sub>2</sub>、5 mM 2-メルカプトエタノール、0.5 mM ATP、25% PEG8000中で、RNAリガーゼ (タカラ) 250ユニット用いて、5'-oligo 1 (5'-UUU GGA UUU GCU GGU GCA GUA CAA CUA GGC UUA AUA CUC GAG UCC GAC -3' / 配列番号：1) 、5'-oligo 2 (5'-UUU CUG CUC GAA UUC CUA ACG AUG UAC GCU CGA GUC CGA C -3' / 配列番号：2) を20°Cで3~16時間連結させた。

#### 【0116】

未連結の5'-オリゴを除去した後、RNaseHフリー逆転写酵素 (Superscript II、ギブコBRL (Gibco BRL) ) でcDNAを合成した。5'末端濃縮cDNAライブラリーを作製するため、ランダムアダプタープライマー (5'-GCG GCT GAA GAC GGC CTA TGT GGC CNN NNN NC-3' / 配列番号：4) 10 pmolを用いて12°Cで1時間インキュベートし、42°Cでさらにもう1時間インキュベートした。

#### 【0117】

第1鎖を合成した後、15 mM NaOH中で65°Cにて1時間インキュベートすることにより、RNAを分解した。オリゴキャップポリ(A)+ RNA 1 mgから作製したcDNAは、XL PCRキット (パーキンエルマー (Perkin-Elmer) ) を使用し、16 pmolの5' (5'ビオチン-GGA TTT GCT GGT GCA GTA CAA CTA GGC TTA ATA-3' / 配列番号：5、または5'ビオチン-CTG CTC GAA T TC AAG CTT CTA ACG ATG TAC G-3' / 配列番号：6) および3' (5'-GCG GCT GAA GAC GGC CTA TGT-3' / 配列番号：7) PCRプライマーにより100 $\mu$ l量で増幅した。ランダムアダプタープライマーで伸張したcDNAについては、増幅サイクルを、94°Cで1分間、58°Cで1分間、72°Cで2分間の10サイクルとした。PCR産物をフェノール：クロロホルム (1:1) で1度抽出してエタノール沈殿し、MmeIタイプIIS制限酵素 (グダニスク大学技術移転センター、ポーランド、グダニスク) で切断した。切断は、300 $\mu$ lの10 mM HEPES、pH 8.0、2.5 mM 酢酸カリウム、5 mM酢酸マグネシウム、2 mM DTT、および40 $\mu$ M S-アデノシルメチオニン中でMmeI 40ユニットを使用し、37°Cで2.5時間行った。

#### 【0118】

切断した5'-末端cDNA断片を、ストレプトアビジンコーティングした磁気ビーズ (ダイナル (Dynal) 、ノルウェー、オスロ) に結合させた。ビーズに結合したcDNA断片を、供給緩衝液中にT4 DNAリガーゼ4ユニットを含む反応液16 $\mu$ l中で、16°Cで2.5時間、互いに直接連結させた。プライマー5'-GGA TTT GCT GGT GCA GTA CAA CTA GGC -3' / 配列番号：8および5'-CTG CTC GAA TTC AAG CTT CTA ACG ATG-3' / 配列番号：9を用いてPCRし、ダイタグを増幅した。PCR産物はポリアクリルアミドゲル電気泳動 (PAGE) で解析し、XhoIで切断した。ダイタグを含むバンドを切り出して自己連結させ、長いコンカテマーを作製した。pZero 1.0 (インビタロジエン) のXhoI部位に、このコンカテマーをクローニングした。M13フォワードおよびM13リバースプライマーを用いて、PCRによりコロニーをスクリーニングした。600 bpを超える挿入断片を含むPCR産物を、Big Dyeターミネーターver. 3でシークエンシングし、3730 ABI自動DNAシークエンサー (アプライドバイオシステムズ、カリフォルニア州) を用いて解析した。電気泳動図はすべて目視検査により再度解析し、あいまいな塩基を調べ誤読を訂正した。SAGE 2000ソフトウェア (バージョン4.12) を用いて、各タグの存在量を定量した。

#### 【0119】

5'SAGEタグの対応遺伝子との関連性

転写開始点の同定における5'SAGEタグの有効性を評価するため、5'SAGEタグを現行のcDNA/ESTデータベースとアラインすることを避けた。その配列が常に転写開始点から読まれているとは限らないからである。代わりに、<http://alps.gi.k.u-tokyo.ac.jp/>で公開されているアライメントプログラムALPSを用いて、我々の5'-タグを<http://genome.ucsc.edu/>で利用可能なヒトゲノム配列、NCBI build 34とアラインすることを試みた。センス方向で一致したタグのみをこの解析で考慮した。

#### 【0120】

次に、Gene Resource Locatorデータベース(Honkura, T., Ogasawara, J., Yamada, T. & Morishita, S. The Gene Resource Locator: gene locus maps for transcriptome analysis. Nucleic Acids Res. 30, 221-225, 2002 URL <http://grl.gi.k.u-tokyo.ac.jp/>)、UniGene (Build 162) (Wheeler, D.L. Database Resources of the National Center for Biotechnology. Nucleic Acids Res. 31, 28-33, 2003 URL <ftp://ftp.ncbi.nih.gov/repository/UniGene/>)等の様々なリソースにおける配列のアライメントのデータベースを利用し、各5'-タグのアライメント位置の近傍を検索して、対応する転写物を見い出した。主だった問題点は、レトロトランスポジションおよびゲノム重複が原因で、1つの5'-タグが、その多くが非コード領域である複数の位置とアラインされてしまうことであった。この問題は、UniGeneデータベースで注釈づけられている遺伝子コード部位を選択することにより解決した。3'-タグは3'-末端エクソンに集まることが多いが、5'-タグは第1エクソンに当たる必要はない。したがって、各5'-タグのアライメント位置から500 bpの距離の範囲内で検索を行った。

#### 【0121】

##### 既知の5'転写開始点との一致

各5'SAGEタグがアラインする位置とその対応する遺伝子間の距離が短いことから、5'-タグは既知の5'転写開始点とほぼ一致することが示唆された。しかし、距離を算出するためには、5'-タグの近傍では、選択的スプライシングが原因で複数のcDNA/EST配列アライメントが頻繁に見られることに留意しなければならない。この状況を解決し距離に固有の値を割り当てるため、5'-タグに最も近いアライメントを選択した。5'-タグが対応するcDNAの上流領域に位置する場合は、距離はマイナスであると定義した。そうでなければ、値はプラスまたはゼロである。特に、距離ゼロとは完全な一致を示す。全体的な距離の分布を見るため、mRNA開始点の-500～+200 ntの5'SAGEタグ出現率の総数を算出した。RefSeq、UniGene (GRL) およびDBTSSデータベースを別々に使用し、転写開始点をカバーする範囲の相違を見た。

#### 【0122】

##### 結果

##### 5' SAGE法

転写開始部位に関して包括的な情報を得るために、本発明者らは、オリゴキヤッピング法を用いて5' SAGEを開発した。5' SAGE法は転写物の5'末端に由来する19～20 bpのタグを生成し、これを迅速に分析してゲノム配列データにマッチさせることができる。図1は、5' SAGE法の戦略を示す。

#### 【0123】

##### ゲノムのマッピング

この方法を用いて、本発明者らは、試験細胞株としてHEK293細胞において発現された転写物25,684個の特徴を調べ、これらをヒトゲノム配列と比較した。全体でタグ19,893個が、異なるタグ13,404個を表すゲノム配列と完全にマッチした(表1)。

異なるタグ13,404個の80% (タグ10,706個) が唯一の位置にマップされた。ゲノムにおいて多数の部位にマッチしたタグは、2つの遺伝子座(loci)にマップされたタグが11.1% (タグ1483個)、3～99の遺伝子座(loci)にマップされたタグが8.1% (タグ1090個)、そして100以上の遺伝子座(loci)にマップされたタグは0.9% (タグ125個) であった。多数のゲノム座にマッピングされたタグは、ほとんどがレトロトランスポゾンエレメント、反復配列、または偽遺伝子に対応する。

【0124】  
【表1】

| Tag loci in genome # | 5'-end SAGE tag to genome# |                                  |                           | 3'-end SAGE tag to genome# |                                  |                           |
|----------------------|----------------------------|----------------------------------|---------------------------|----------------------------|----------------------------------|---------------------------|
|                      | Tags mapped to genome (%)  | Unique Tags mapped to genome (%) | Relative expression level | Tags mapped to genome (%)  | Unique Tags mapped to genome (%) | Relative expression level |
| 1 loci/genome        | 15,448 (77.7)              | 10,706 (79.9)                    | 1.44                      | 34,139 (63.2)              | 11,613 (75.3)                    | 2.94                      |
| 2 loci/genome        | 2,037 (10.2)               | 1,483 (11.1)                     | 1.37                      | 6,739 (12.5)               | 1,395 (9.0)                      | 4.83                      |
| 3~99 loci/genome     | 2,275 (11.4)               | 1,090 (8.1)                      | 2.09                      | 12,265 (22.7)              | 2,039 (13.2)                     | 6.02                      |
| >100 loci/genome     | 133 (0.7)                  | 125 (0.9)                        | 1.06                      | 907 (1.7)                  | 376 (2.4)                        | 2.42                      |
| Total tag            | 19,893 (100)               | 13,404 (100)                     | 1.40                      | 54,050 (100)               | 15,422 (100)                     | 2.13                      |

表1. SAGEタグとゲノムとの実験的な照合

5'-end SAGE tag to genome:ゲノムにマップされた5'SAGEのタグの数

3'-end SAGE tag to genome:ゲノムにマップされた3'SAGEのタグの数

Tags mapped to genome(%):ゲノムにマップされたタグの数(%)

Unique Tags mapped to genome(%):ゲノムにマップされたユニークなタグの数(%)

## Relative expression level:相対発現レベル

#: 18 bp 5' SAGEタグを用いてゲノムにヒットしたタグの数。マッピングは材料と方法の章に記述したとおりに実施した。ゲノムにヒットしなかったタグは、シークエンシングしたタグ25,684個中5,791個であった。相対的発現レベルは、ライブラリにおいて認められた転写物タグの総数を異なるタグの数によって除することによって決定した。

##: 20 bp 3' SAGEタグを用いてゲノムにヒットしたタグの数。マッピングは材料と方法の章に記述したとおりに実施した。ゲノムにヒットしなかったタグは、シークエンシングしたタグ81,211個中27,162個であった。

## 【0125】

## mRNA開始部位へのマッピング

次に、本発明者らは、5' SAGEタグがmRNA開始部位にマッチするか否かを推定した。本発明者らは、参考配列データベース (RefSeq)、調節領域におけるシス要素およびオルタナティブスプライシング転写物に関する情報を含む遺伝子マップを構築するGene Resource Locator (GRL)、およびヒト完全長cDNAsの系統的な5'末端配列を含む DataBase of Transcriptional Start Site (DBTSS) (Suzuki, Y. et al. DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. Nucleic Acids Res 30, 328-331, 2002) を含む三つのデータベースを用いた。図2は、距離の分布を示し、表2は、距離が短いタグの発生比率を示し、本発明者らの5' SAGEタグがそれぞれのデータベースの開始部位情報と十分に一致することを示している。それぞれのデータベースにマッピングされたタグの85.8%~98.2%が、mRNA開始部位の-500ヌクレオチド~+200ヌクレオチドにマップされた。

## 【0126】

特に、5' SAGEタグの23.5~49.3%が、これらのデータベースにおいて定義された転写開始部位 (TSS) の上流の領域にヒットする。その上、本発明者らは、5' SAGEタグによるTSSでのヌクレオチド選択性 (nucleotide preference) を調べた。TSSのヌクレオチドは、ヒト遺伝子276個におけるmRNA 5880個を用いて、A (47%)、G (28%)、C (14%)、およびT (12%) であると報告されている (Suzuki, Y. et al. Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites. EMBO Rep 2, 388-393, 2001)。本発明者らのデータはまた、最初のヌクレオチドの使用に関して非常に類似の百分率を示した：A (41%)、G (32%)、C (17%)、およびT (10%)。併せて考慮すると、本発明者らの5' SAGEタグ法は、TSSを正確に同定することができる。データは本発明者らに、正確な転写開始部位情報を提供するのみならず、プロモーターの利用を分析するためのリソースを提供する。興味深いことに、ゲノムにマッチしないタグは、本研究において5' SAGEにおける全シークエンシングタグの33%を占めた。それらの中で、ゲノムにマッチしない5' SAGEタグの第一のヌクレオチドの39%もまたAであった。ゲノムにマッチしなかったタグのいくつかは、ゲノムにおける単一のヌクレオチド変異または欠失を有する領域にヒットすると見なすことができる。

## 【0127】

【表2】

| Distance from start site<br>of each database (nt) | Tag number (%)      |                     |                     |
|---|---------------------|---------------------|---------------------|
|   | RefSeq              | UniGene (GRL)       | DBTSS               |
| -500 ~ -201                                       | 349 (3.2)           | 204 (1.5)           | 160 (1.6)           |
| -200 ~ -51  | 887 (8.1)           | 335 (2.4)           | 253 (2.5)           |
| -50 ~ -1  | 4,179 (38.1)        | 3,957 (28.8)        | 1,965 (19.5)        |
| 0 ~ +50   | 3,173 (28.9)        | 8,673 (63.2)        | 7,149 (70.8)        |
| +51 ~ +200  | 837 (7.6)           | 311 (2.3)           | 209 (2.1)           |
| (-500 ~ +200)                                     | 9,425 (85.8)        | 13,480 (98.2)       | 9,736 (96.4)        |
| <b>Total tags</b>                                 | <b>10,982 (100)</b> | <b>13,723 (100)</b> | <b>10,098 (100)</b> |

表2. 各データベースにおけるmRNAの開始部位と対応する5'SAGEタグとの距離

Distance from start site of each database(nt)：各データベースの開始点からの距離  
(スクレオチド)

各データベースにおける遺伝子の5'末端へのマッピングにおいて一致するタグを図2に示したように分析した。

### 【0128】

新規遺伝子またはアノテーションされていない遺伝子の同定

特徴がわかつていない遺伝子を同定するために、5'SAGEタグをゲノム配列、RefSeq、およびESTデータベースと比較した。ゲノムにおいて単一の座を有するユニークなタグ10,706個のうち、タグ9,376個を、その対応するUniGene ESTに関連させることができる(表3)。さらに、5'SAGEのユニークなタグ6,418個は、DBTSSにおける既知の遺伝子に関連していた。残りのタグ(12.4%)は、既知の遺伝子のイントロン内の領域(5.4%)または特徴がわかつていない領域(6.6%)にマッチした。特徴がわかつていない領域にマッチしたタグは、主に二つの部位にヒットした；

- (1) 全く特徴がわかつていない領域、
- (2) 特徴がわかつていないESTの領域

そのような遺伝子の発現に関する証拠があれば、3'SAGEを参考にすることによって完全長の形の新規遺伝子を発見するために役立つはずである。

### 【0129】

【表3】

| Gene/exon category            | UniqueTags mapped to genome (tags occurrences ) |                        |
|-------------------------------|---|------------------------|
|                               | 5'SAGE  | 3' SAGE                |
| <b>Previously annotated</b>   |   |                        |
| Known genes                   | 9,376 (13,674)                                  | 8,359 (27,996)         |
| <b>Previously unannotated</b> |   |                        |
| Internal exons (Intron)       | 515 (713)                                       | 1,329 (2,442)          |
| genome                        | 815 (1,061)                                     | 1,925 (3,701)          |
| <b>Total</b>                  | <b>10,706 (15,448)</b>                          | <b>11,613 (34,139)</b> |

表3：特徴づけされていない候補遺伝子とエキソンの同定

gene/exoncategory：遺伝子／エキソンの分類

Unique tags mapped to genome(tags occurrences)：ゲノムにマッピングされたユニークなタグ（タグの出現頻度）

Previously annotated：既にアノテーション済み

Known gene：既知遺伝子

Previously unannotated：未だアノテーションされていない

Internal exon(Intron)：内部エキソン（イントロン）

genome：ゲノム

total：総数

10,706個がユニークな位置にマップされ、9,376個が対応するUniGene ESTに関連付けられる。

### 【0130】

SAGEは、転写物の量に基づく定量的情報を得るために用いることができる非常に強力な方法である。表4は、HEK 293細胞における転写物プロファイルの5'末端を示す。最も発現量の多い遺伝子は、neurofilament 3(NEF3)として同定され、その発現頻度は1.43%であり、これに次いで複数座にヒットした遺伝子、および elongation factor 2であった。NEF3、heat shock 70kDa protein 1A (熱ショック70 kDa蛋白質1A)、calreticulin (カルレチキュリン)、およびheterogeneous nuclear ribonucleoprotein H1 (異種核リボ核蛋白質H1) のようないくつかの遺伝子は、異なるタグを示した。いくつかの遺伝子は、異なるTSSから転写されたことが示唆される。例えば、heat shock 70kDa protein 1Aは、異なる転写開始部位8個から転写される。calreticulinは、異なる転写開始部位7個から転写される。これらの結果は、個々の転写開始部位が遺伝子発現に関連するであろうことを示唆している。

### 【0131】

【表4】

| Tag sequence         | Tag count | Related Unigene cluster     | Related refseq                | Gene   |
|----------------------|-----------|-----------------------------|-------------------------------|--|
| GCTCTGACGCCACCGC     | 286       | Hs_71346                    | NM_005382                     | Homo sapiens neurofilament 3 (150kDa medium) (NEF3), mRNA                        |
| CTTTCCTCTTCTAACTAACG | 171       | Hs_75309                    | NM_001961                     | Multiple hit to genome   |
| CTCTCCGCGTCGTC       | 120       |                             |                               | Multiple hit to genome   |
| TACCTGTGTGATCCCTCC   | 117       |                             |                               | ectaryotic translation elongation factor 2                                       |
| CTGTGATCCCTCCAGT     | 89        | Hs_274402,Hs_75452,Hs_80288 | NM_005345,NM_005346           | Multiple hit to genome   |
| AACGCTTACGCTTGGAGAG  | 83        |                             | NM_031243,NM_002137           | heat shock 70kDa protein 1A  |
| AGTACAGAGCGCGGG      | 75        | Hs_232400                   |                               | heterogeneous nuclear ribonucleoprotein A2/B1                                    |
| CTCATTTACTCGCTCTC    | 68        |                             |                               | Multiple hit to genome   |
| GTGACAGCCACCCGCC     | 65        | Hs_71346                    | NM_005382                     | Homo sapiens neurofilament 3 (150kDa medium) (NEF3), mRNA                        |
| AFTTGATGATGATGCGG    | 57        |                             |                               | Multiple hit to genome   |
| ACGCTTGTACGCCACAC    | 56        | Hs_71346                    | NM_005382                     | Homo sapiens neurofilament 3 (150kDa medium) (NEF3), mRNA                        |
| CTTITTCGCAACGGTTT    | 55        |                             |                               | Multiple hit to genome   |
| AATTCTACGCGACCCG     | 54        | Hs_446528                   | NM_001007                     | Multiple hit to genome   |
| ACCGCGAGACCCGCTCC    | 53        | Hs_426930,Hs_510444         | NM_001101                     | ribosomal protein S4, X-linked   |
| CTTCACTGAAAGCGCGC    | 52        | Hs_5652,Hs_569234           | NM_006998                     | actin, beta  |
| ATATTCCTTACTCTCTG    | 48        | Hs_14376,Hs_500737          |                               | glutamine nucleotide binding protein (G protein), beta polypeptide 2-like 1      |
| CTCAGTCGCGCTCCG      | 43        |                             |                               | Homo sapiens X (inactive)-specific transcript (XIST) on chromosome X             |
| ATTCTTGTAAAGCGGAC    | 42        | Hs_268849                   | NM_006708                     | actin, gamma 1.  |
| CCTCCCTCATCACAGCGC   | 37        | Hs_15589                    | NM_004744                     | glyoxalase 1   |
| CTTCTCTACTCTCTCG     | 37        |                             |                               | Multiple hit to genome   |
| AGTACAGCTCGGCC       | 35        | Hs_402752                   | NM_003487,NM_139215           | PPAR binding protein   |
| CACCTTGTGAGCTGTC     | 34        | Hs_146550                   | NM_006998                     | Multiple hit to genome   |
| CTCTCTCTACTCGCAAG    | 33        | Hs_5662,Hs_509234           | NM_005348                     | TAF15 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 68kDa |
| AGTCTGCTGCTGCC       | 30        | Hs_446579,Hs_449534         | NM_006346                     | myosin, heavy polypeptide 9, non-muscle  |
| AGTCTCGCGCTGACGC     | 29        | Hs_211602                   | NM_004343                     | granine nucleotide binding protein (G protein), beta polypeptide 2-like 1        |
| GTCCCTACTCGAGCGG     | 29        | Hs_353170                   |                               | heat shock 90kDa protein 1, alpha  |
| AAAACGCCAGCCTCTAGG   | 27        | Hs_75452                    |                               | SMC1 structural maintenance of chromosomes 1-like 1 (yeast)                      |
| ATTTCGCTTACGCCACG    | 26        | Hs_202166                   |                               | calreticulin   |
| AGGCATTTGAGCAGCCAG   | 25        |                             |                               | FLJ36598   |
| AGTGGCGGAGCGCGG      | 25        | Hs_192374                   | NM_003299                     | heterogeneous nuclear ribonucleoprotein H1 (H)                                   |
| CCCAATTCTACTGCGCAC   | 25        | Hs_446528                   | NM_001007                     | tumor rejection antigen (gp90) 1   |
| CTCGTGGCAGTAGTCG     | 25        | Hs_380118,Hs_460941         |                               | ribosomal protein S4, X-linked   |
| GTCGCTGAGCGCTGCCG    | 25        | Hs_2795                     | NM_005566                     | RNA binding motif protein, X-linked  |
| CATTCGCTTACGCCAGG    | 24        | Hs_202166                   |                               | lactate dehydrogenase A  |
| CTCTTCTCTTACGCCAGG   | 24        |                             |                               | heterogeneous nuclear ribonucleoprotein H1 (H)                                   |
| GACTATTTGTCGGCGC     | 24        | Hs_280311                   | NM_003299                     | Multiple hit to genome   |
| ACCTCTTCTTACCTTAC    | 23        | Hs_279806                   | NM_004396                     | myosin, heavy polypeptide 10, non-muscle   |
| CTTCTCTCCGCTGACG     | 23        |                             |                               | DEAD (Asp-Glu-Ala-Asp) box polypeptide 5   |
| AGTATCTGTTGGTACCCG   | 22        | Hs_333455,Hs_311035         | NM_001428                     | Multiple hit to genome   |
| CAATTCTTAAGCGCCACG   | 22        | Hs_446628                   | NM_006158                     | endothelial 1, (alpha)   |
| GCACACAGCCATCCATCC   | 22        | Hs_107600                   | NM_004741                     | ribosomal protein S4, X-linked   |
| AGTGAGCGCTTATTCG     | 19        | Hs_75337,Hs_467172          | NM_004741                     | nucleolar and coiled-body phosphoprotein 1                                       |
| CTCTTCCAGCCAGCGCC    | 19        |                             |                               | neurofilament, light polypeptide 68kDa   |
| CITTCCTCACCGCTCC     | 19        | Hs_374596                   | NM_003295                     | endothelial 1, (alpha)   |
| GGGTCTTCTCTGCTCTG    | 19        | Hs_180909                   | NM_181595,NM_181597,NM_002574 | tumor protein, translationally-controlled 1                                      |
| ATATAAGGCTGGGGG      | 18        | Hs_427152                   |                               | peroxiredoxin 1  |
| CTTITGGCTCTBACCC     | 18        | Hs_78996,Hs_449476          | NM_002592                     | high density lipoprotein binding protein (vigin)                                 |
|                      | 18        |                             |                               | proliferating cell nuclear antigen   |
|                      |           |                             |                               | Multiple hit to genome   |

表4. HEK293細胞における転写プロファイルの5'末端

Tag sequence：タグ配列

Tag count：タグの計数値

Related Unigene cluster：関連Unigeneクラスター

Related refseq：関連 refseq

Gene：遺伝子

HEK293細胞において発現していた上位50の5'末端転写物をリストした。タグ配列は18-bpのSAGEタグを示す。タグとそれに対応するUnigene/ESTを示した。

【0132】

5' と 3' SAGEタグ発現の一致

本発明者らはまた、5' SAGEの精度を確認するために、同じ細胞においてmRNAの3'-Long SAGEを試みた。3'-Long SAGEにおいて、本発明者らは、HEK293細胞株において発現され

た転写物タグ81,212個の特徴を調べた。全体でタグ54,050個が、異なるタグ15,423個を表すゲノム配列にマッチした（表1）。異なるタグ15,423個の75%（タグ11,613個）がゲノムにおいて一つの部位にマッチした。さらに、3' SAGEタグ8,359個がUniGene ESTにおける既知の遺伝子に関連した（表3）。ゲノムにおいて多数の部位にマッチしたタグは、2つの座にマッチしたタグが9%（タグ1395個）、3～99の座にマッチしたタグが13.2%（タグ2,039個）、および100以上の座にマッチしたタグが2.4%（タグ376個）であった。ゲノムにおいて多数の部位にマッチしたタグの割合は、5' SAGEと3' SAGEのあいだでよく似ていた（表2）。一方、5' SAGEタグは、3' SAGEタグと比較して非常に不均一であった。

### 【0133】

Sehaらも同様に、ゲノム1個あたり10個より多いコピーを示すタグは、ゲノム1個あたりコピー1個のみを示すタグより、平均して高度に発現されることを示した（Saha, S. et al. Nat Biotechnol 20, 508-512, 2002）。本発明者らのデータはまた、3～99座/ゲノムでは、5' SAGEおよび3' SAGEライプラリにおける他の分画より相対的な発現レベルが高いことを証明した。これはレトロトランスポジターを通しての遺伝子発現と遺伝子複製との相関のメカニズムによる。二つのライプラリのあいだの類似性の程度を推定するために、5' SAGEと3'-Long SAGEのあいだで発現された遺伝子を比較した。

### 【0134】

5'および3'タグは5'末端および3'末端から個々に無作為に採取するため、5'タグが特定の完全長のcDNA配列に関連する確率は、3'タグがcDNAにマッチする確率と一致すると予想される。しかし、完全長のcDNA配列またはオルタナティブスプライシング転写物のコレクションが不完全であるために、たとえこれらのタグが同じコード領域に由来するとしても、5'タグと3'タグのあいだに正確な一致を決定することは簡単ではない。一つの有望なアプローチは、エキソンを共有するESTアラインメントをまとめて遺伝子コード座のようなクラスターとして扱い、5'および3' SAGEタグをこれらのクラスターおよびその上流の領域にマッピングして、5'および3' SAGEタグ発現のあいだの一一致を発見することであろう。このようにして、本発明者らはそれぞれの遺伝子コード領域に関する3'および5'タグの対の発生数を計数し、そして図3の二次元平面に全ての対を表した。発現パターンの比較によって、ほとんどの遺伝子が双方のライプラリで類似のレベルで発現されることが判明した。しかし、いくつかの転写物は、有意に異なるレベルで発現され、5' SAGEと3' SAGEライプラリのピアソン相関係数は0.36で、中等度の類似性を示した。

### 【0135】

相関が中等度であった理由は、5' SAGEと3' SAGEライプラリからの頻度の分散による。以下に由来する配列のように、これらのタグの出現に関するいくつかの可能性がある。

- (1) 5' SAGEおよび3' SAGEにおけるPCR増幅の誤差
- (2) 3' SAGEにおいてNlaIII制限部位を占有すると予想されるであろう少数の遺伝子
- (3) 5' SAGEにおいてXhoI制限部位を占有すると予想されるであろう少数の遺伝子
- (4) 5' SAGEおよび3' SAGEにおけるmRNAの未知のスプライシング変種
- (5) 多数のゲノム座に対するタグのヒットに関する注釈誤差、またはゲノムへのEST注釈誤差

### 【0136】

本研究は、例としてHEK 293細胞において発現された遺伝子の分画のみを同定した。発現された遺伝子の大要を詳細に記述するためには、多様な異なる細胞タイプおよび環境条件からのかなり多数のタグが必要であろう。データが蓄積されれば、5' と 3' SAGEタグの発現の一一致に関する問題が解決される可能性がある。

### 【0137】

#### 考察

mRNA開始部位（Suzuki, Y. et al. Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites. EMBO Rep 2, 388-393, 2001）とポリアデニル化切断部位（Pauws, E., van Kampen, A.H., van de Graaf, S.A., de Vijlder, J.

J. & Ris-Stalpers, C. Heterogeneity in polyadenylation cleavage sites in mammalian mRNA sequences: implications for SAGE analysis. Nucleic Acids Res 29, 1690-1694, 2001)が不均一性を示すことは、いくつかの研究グループによって報告された。Shirakiらは構築の際の特定の遺伝子のTSSの差を報告したが(Shiraki, T. et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci U S A 100, 15776-15781, 2003)、本発明者らのデータは、TSSの多様性が細胞に既に存在することを示している。その上、本発明者らのデータは、5' SAGEおよび3' SAGE法によってTSSと3'末端領域の不均一性に関する直接の証拠を提供する。

#### 【0138】

例えば、PPAR結合蛋白質はTSS 1個と3' SAGEタグ部位2個とを有し、リボソーム蛋白質S4はTSS 16個と3' SAGEタグ部位1個とを有し、カルレチキュリンはTSS 7個と3' SAGEタグ部位1個とを有する。さらに、オルタナティブmRNAスプライシングは、ヒトプロテオームの複雑性に極めて重要に関与している。最近のゲノム研究から、ヒト遺伝子の40~60%がオルタナティブスプライシングされていることが証明されている(Modrek, B. & Lee, C. A genomic view of alternative splicing. Nature Genetics 30, 13-19, 2002)。点突然変異の15%がmRNAスプライシング欠損によってヒト遺伝疾患を引き起こすと推定されている(Krawczak, M., Reiss, J. & Cooper, D.N. The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. Hum Genet 90, 41-54, 1992)。

#### 【0139】

Zavolanらは、多数のスプライス型を有する転写ユニットは、その49%が、オルタナティブ転写開始の使用が最初のエキソンのオルタナティブスプライシングを伴う転写物を含むことを報告した(Zavolan, M. et al. Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. Genome Res 13, 1290-1300, 2003)。本発明者らはまた、ペルオキシレドキシン4 (NM\_006406) のようないくつかの遺伝子のそれぞれのmRNA開始部位が、mRNAの異なるスプライシングバリエントを表すのみならず、遺伝子発現の異なる量を表すことを発見した。このことは、オルタナティブトランスクリプションがオルタナティブスプライシングをしばしば誘導する可能性があることを意味する。

#### 【0140】

最近、キャップトラッパーシステムを用いて転写開始点を同定する新規方法が報告されている(Shiraki, T. et al. Proc Natl Acad Sci U S A 100, 15776-15781, 2003)。しかし、mRNA開始部位のマッピングの同定効率は、それらの方法では顕著ではない。本発明者らの研究において記述された5' SAGE法によって、TSSを正確にマッピングでき、同様に遺伝子発現の頻度を確立することができる。

#### 【0141】

結論すると、5' SAGE法を利用すれば、ゲノムの注釈がかなり促進される可能性がある。5' SAGEは、遺伝子配列のアприオリ知識に依存しない少ない高スループット発見アプローチの一つであるため、そのようなデータによって、in silico遺伝子予測の独立した確認および未注釈領域の同定が同時に可能となるであろう。さらに、5' SAGE法は、5' UTR/プロモーター領域においてSNPを発見するために有用であろう。異なるタイプの特異的mRNA開始部位から転写された遺伝子を包括的に同定すれば、ヒトゲノムの機能的複雑性に対する説明のみならず、癌、免疫、および神経疾患のような様々な障害の診断の基礎に対して新しい洞察が得られる。

最後に、5'末端の多様性を考慮に入れると、遺伝子発現の頻度を決定するためには、3' SAGEより5' SAGEを行うことがより適当であろう。

#### 【産業上の利用可能性】

#### 【0142】

本発明は、遺伝子タグの取得に有用である。遺伝子タグは、遺伝子に固有の塩基配列情

報である。したがって、ある遺伝子ライブラリーにおけるタグの出現頻度は、そのライブラリーを構成する全ての遺伝子の発現状態を反映していると考えられる。そのため、遺伝子タグは、遺伝子発現解析に有用である。特に本発明によって得ることができる遺伝子タグは、全てのmRNAが有している5'末端の構造に基づいて生成される。したがって、本発明によって生成されるタグに基づく遺伝子発現解析の結果は、より信頼性が高い。

#### 【0143】

また本発明のタグは、mRNAの5'末端領域の塩基配列情報を含んでいる。したがって、本発明によって生成されるタグの塩基配列情報に基づいて、ゲノムにおける転写開始点を同定することができる。また本発明のタグの塩基配列情報に基づいてデザインされたオリゴヌクレオチドは、全長cDNAの合成用プライマーとして利用することができる。

#### 【図面の簡単な説明】

#### 【0144】

【図1】本発明に基づく、遺伝子タグの取得方法の例を示す図。mRNAを半分に分けて、mRNAのCap構造を、II<sub>s</sub>型制限エンドヌクレアーゼであるMmeIおよびXhoI制限酵素部位を含む二つのタイプの合成オリゴヌクレオチドに酵素的に置換した。次に、オリゴキャッピングmRNAをdTアダプタープライマーによってcDNAの第1鎖に変換した。第2鎖を、PCRを用いて、ビオチン結合5'プライマーおよびdTアダプタープライマーによって合成した。二本鎖cDNAを、その認識部位から20 bp離れたところで切断するMmeIによって切断した。ストレプトアビジンビーズに結合させることによって5' cDNAを単離した後、タグの二つのプールを互いにライゲーションした。

【図2】UniGeneおよびDBTSS配列におけるmRNA開始部位と比較した5' SAGEタグの距離。距離は、上流（-）および下流（+）のヌクレオチド（x-軸）の数として示す。UniGeneにおけるmRNA開始部位を0として示す。5' SAGEタグの頻度をy-軸に示す。それぞれの5' SAGEタグとその対応する遺伝子とを配置した位置の距離が短ければ、5'タグが既知の5'転写開始部位とほぼ一致することを意味している。本発明者らは、転写開始部位のその範囲の差を調べるために、UniGeneおよびDBTSSデータベースを別々に用いた。

【図3】5' SAGEタグと3' SAGEタグの頻度のスキャッタープロット。5' SAGEおよび3' SAGEからゲノムにおける一つの座にヒットしたタグを、実施例2の材料と方法の章に記載するように分析した。この図において、双方の軸は対数で表記した。

## 【配列表】

## SEQUENCE LISTING

<110> Post Genome Institute Co., Ltd.

<120> Method for producing gene tags

<130> PGI-A0301Y1

<150> JP 2003-402306

<151> 2003-12-01

<160> 43

<170> PatentIn version 3.1

<210> 1

<211> 48

<212> RNA

<213> Artificial

<220>

<223> an artificially synthesized RNA linker sequence

<400> 1

uuuggauuug cuggugcagu acaacuaggc uuauauacucg aguccgac

48

<210> 2

<211> 46

<212> RNA

<213> Artificial

<220>

<223> an artificially synthesized RNA linker sequence

<400> 2

uuucugcucg aauucaagcu ucuaacgaug uacgcucgag uccgac

46

<210> 3

<211> 42

<212> DNA

<213> Artificial

<220>

<223> an artificially synthesized primer sequence

<400> 3

gcggctgaag acggcctatg tggcctttt tttttttt tt

42

<210> 4  
<211> 32  
<212> DNA  
<213> Artificial

<220>  
<223> an artificially synthesized primer sequence

<220>  
<221> misc\_feature  
<222> (26)..(31)  
<223> "n"=a, t, g or c

<400> 4  
gcggctgaag acggcctatg tggccnnnn nc

32

<210> 5  
<211> 33  
<212> DNA  
<213> Artificial

<220>  
<223> an artificially synthesized primer sequence

<220>  
<221> misc\_feature  
<222> (1)..(1)  
<223> Label biotin

<400> 5  
ggatttgctg gtgcagtaca actaggctta ata

33

<210> 6  
<211> 31  
<212> DNA  
<213> Artificial

<220>  
<223> an artificially synthesized primer sequence

<220>  
<221> misc\_feature  
<222> (1)..(1)  
<223> Label biotin

<400> 6

ctgctcgaat tcaagcttct aacgatgtac g

31

<210> 7  
 <211> 21  
 <212> DNA  
 <213> Artificial

<220>  
 <223> an artificially synthesized primer sequence

<400> 7  
 gcggctgaag acggcctatg t

21

<210> 8  
 <211> 27  
 <212> DNA  
 <213> Artificial

<220>  
 <223> an artificially synthesized primer sequence

<400> 8  
 ggatttgctg gtgcagtaca actaggc

27

<210> 9  
 <211> 27  
 <212> DNA  
 <213> Artificial

<220>  
 <223> an artificially synthesized primer sequence

<400> 9  
 ctgctcgaat tcaagcttct aacgatg

27

<210> 10  
 <211> 18  
 <212> DNA  
 <213> Homo sapiens

<400> 10  
 acatctgacc tcatggag

18

<210> 11  
 <211> 18

<212> DNA

<213> Homo sapiens

<400> 11

ctcttcctt gcctaacg

18

<210> 12

<211> 18

<212> DNA

<213> Homo sapiens

<400> 12

tacctggttg atcctgcc

18

<210> 13

<211> 18

<212> DNA

<213> Homo sapiens

<400> 13

ctttcctgt ggcagcag

18

<210> 14

<211> 18

<212> DNA

<213> Homo sapiens

<400> 14

ctcttccgcc gtcgtcgc

18

<210> 15

<211> 18

<212> DNA

<213> Homo sapiens

<400> 15

ctcattgaac tcgcctgc

18

<210> 16

<211> 18

<212> DNA

<213> Homo sapiens

<400> 16

ctggttgatc ctgccagt

18

<210> 17  
<211> 18  
<212> DNA  
<213> Homo sapiens

<400> 17  
ctcagtcgcc gctgccag

18

<210> 18  
<211> 18  
<212> DNA  
<213> Homo sapiens

<400> 18  
cttcactgc aaggcggc

18

<210> 19  
<211> 18  
<212> DNA  
<213> Homo sapiens

<400> 19  
acgctgtgac agccacac

18

<210> 20  
<211> 18  
<212> DNA  
<213> Homo sapiens

<400> 20  
gtgacagcca caccccc

18

<210> 21  
<211> 18  
<212> DNA  
<213> Homo sapiens

<400> 21  
aacggctagc ctgaggag

18

<210> 22  
<211> 18  
<212> DNA

<213> Homo sapiens

<400> 22  
agttagcagca gcgccggg

18

<210> 23  
<211> 18  
<212> DNA  
<213> Homo sapiens

<400> 23  
attcctagtt aaggcggc

18

<210> 24  
<211> 18  
<212> DNA  
<213> Homo sapiens

<400> 24  
aattgtgttc gcagccgc

18

<210> 25  
<211> 18  
<212> DNA  
<213> Homo sapiens

<400> 25  
atatttctta ctctctcg

18

<210> 26  
<211> 18  
<212> DNA  
<213> Homo sapiens

<400> 26  
ctcagtcgcc gctgccaa

18

<210> 27  
<211> 18  
<212> DNA  
<213> Homo sapiens

<400> 27  
aaaacggcca gcctgagg

18

<210> 28  
 <211> 18  
 <212> DNA  
 <213> Homo sapiens

<400> 28  
 ctctttca ctgcaagg 18

<210> 29  
 <211> 18  
 <212> DNA  
 <213> Homo sapiens

<400> 29  
 aatttctacg cgcacccgg 18

<210> 30  
 <211> 18  
 <212> DNA  
 <213> Homo sapiens

<400> 30  
 accgcccaga ccgcgtcc 18

<210> 31  
 <211> 18  
 <212> DNA  
 <213> Homo sapiens

<400> 31  
 agacgcagag tagattgt 18

<210> 32  
 <211> 18  
 <212> DNA  
 <213> Homo sapiens

<400> 32  
 agttcgatcg gtagcggg 18

<210> 33  
 <211> 18  
 <212> DNA  
 <213> Homo sapiens

<400> 33  
agttctcggg cgtacggc 18

<210> 34  
<211> 18  
<212> DNA  
<213> Homo sapiens

<400> 34  
agtgcgttca gcgtcccg 18

<210> 35  
<211> 18  
<212> DNA  
<213> Homo sapiens

<400> 35  
attaaacggt tgcaggcg 18

<210> 36  
<211> 18  
<212> DNA  
<213> Homo sapiens

<400> 36  
ccggccgggg ggcgggcg 18

<210> 37  
<211> 18  
<212> DNA  
<213> Homo sapiens

<400> 37  
cctttggct ctctgacc 18

<210> 38  
<211> 18  
<212> DNA  
<213> Homo sapiens

<400> 38  
ctcagtacag ctccggcc 18

<210> 39  
 <211> 18  
 <212> DNA  
 <213> Homo sapiens

<400> 39  
 ctcttcggc cgcgctgg 18

<210> 40  
 <211> 45  
 <212> DNA  
 <213> Artificial

<220>  
 <223> an artificially synthesized DNA linker sequence

<400> 40  
 tttggatttg ctggcagt acaacttaggc ttaatatccg acatg 45

<210> 41  
 <211> 38  
 <212> DNA  
 <213> Artificial

<220>  
 <223> an artificially synthesized DNA linker sequence

<220>  
 <221> misc\_feature  
 <222> (38)..(38)  
 <223> C7-amino-modified

<400> 41  
 tcggatatta agcctagttg tactgcacca gcaaattcc 38

<210> 42  
 <211> 43  
 <212> DNA  
 <213> Artificial

<220>  
 <223> an artificially synthesized DNA linker sequence

<400> 42  
 ttctgctcg aattcaagct tctaacgatg tacgtccgac atg 43

<210> 43  
<211> 36  
<212> DNA  
<213> Artificial

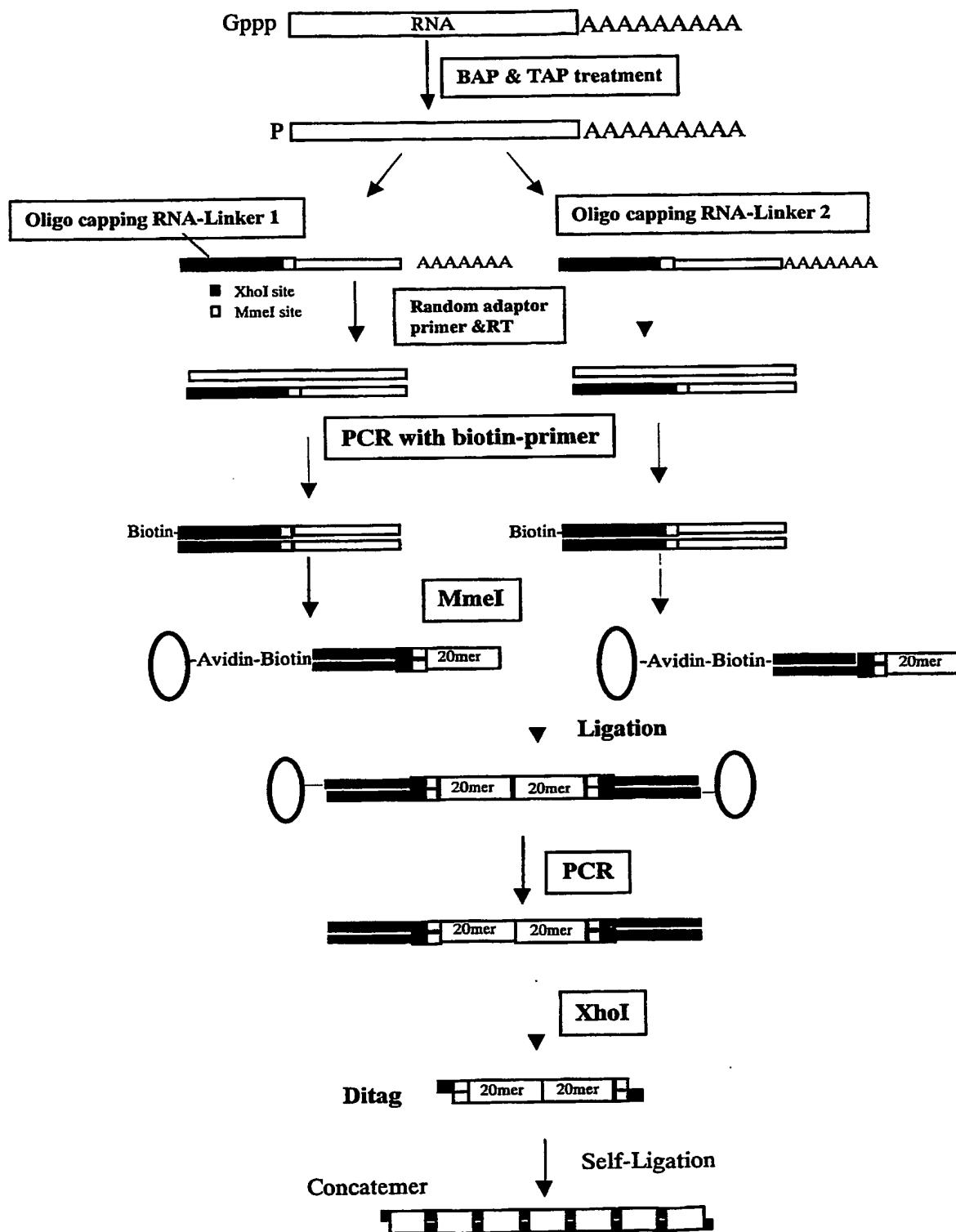
<220>  
<223> an artificially synthesized DNA linker sequence

<220>  
<221> misc\_feature  
<222> (36)..(36)  
<223> C7-amino-modified

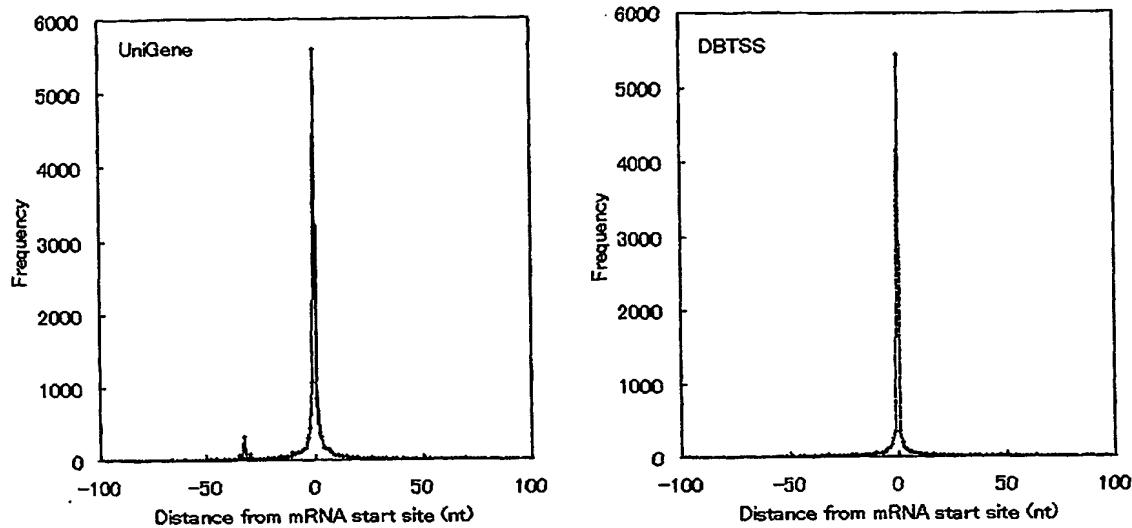
<400> 43  
tcggacgtac atcgtagaa gcttgaattc gagcag

36

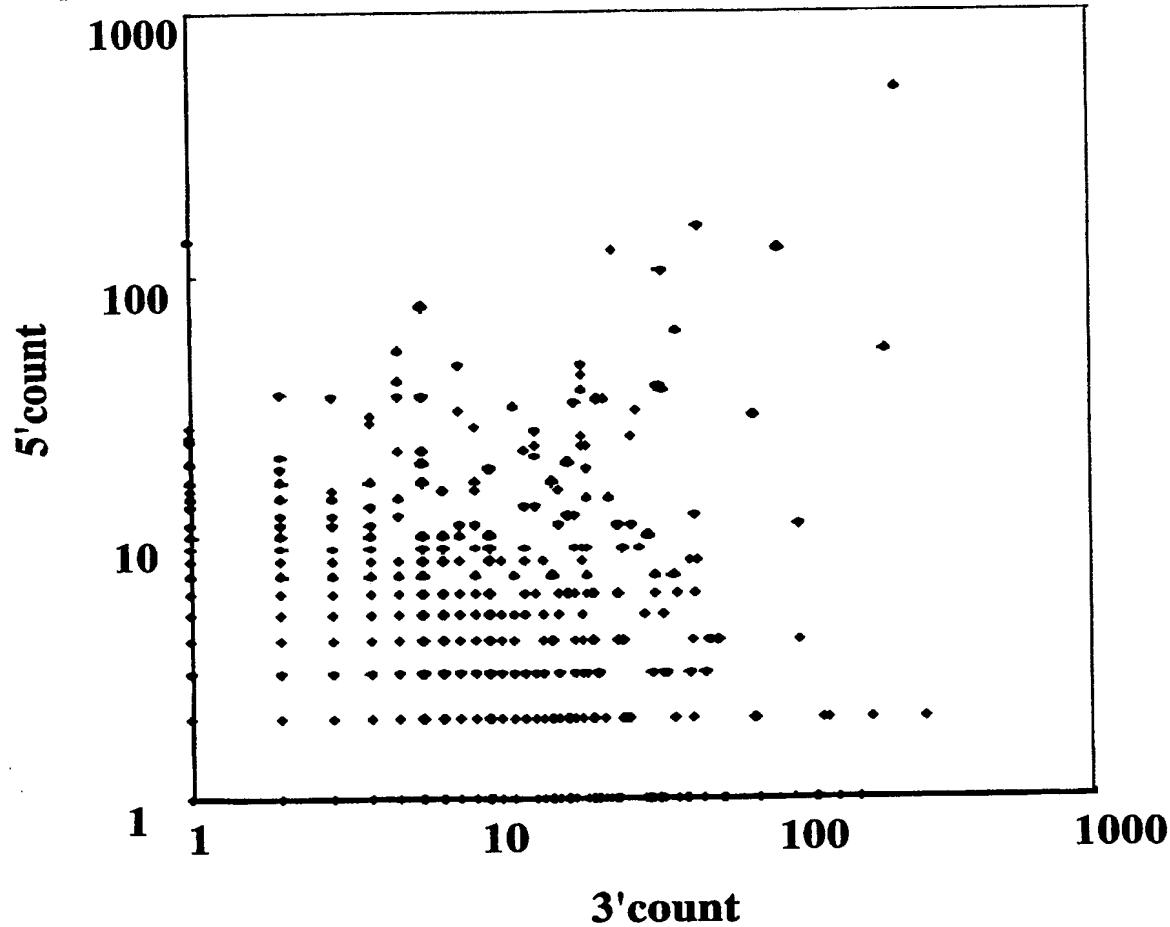
【書類名】図面  
【図1】



【図2】



【図3】



【書類名】要約書

【要約】

【課題】本発明の課題は、遺伝子タグの生成方法の提供である。

【解決手段】mRNAの5'末端の塩基配列をタグとして生成するための方法が提供された。本発明の方法は、CAP構造にII<sub>s</sub>型制限酵素の認識配列を含むII<sub>s</sub>リンカーを連結したmRNAを鋸型としてcDNAを合成する工程を含む。このcDNAにII<sub>s</sub>型制限酵素を作用させることによって、mRNAの5'末端の塩基配列からなるタグが生成される。

塩基配列に依存せず、あらゆるmRNAからタグを生成することができる。本発明のタグの塩基配列情報に基づいて、転写開始点の同定方法や、全長cDNA合成用プライマーが提供される。

【選択図】図1

特願 2004-006630

出願人履歴情報

識別番号 [501005184]

1. 変更年月日 2003年12月 5日

[変更理由]

住所変更

東京都文京区本郷3-38-1 本郷イシワタビル6F  
株式会社ポストゲノム研究所

**This Page is Inserted by IFW Indexing and Scanning  
Operations and is not part of the Official Record.**

## **BEST AVAILABLE IMAGES**

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

**BLACK BORDERS**

**IMAGE CUT OFF AT TOP, BOTTOM OR SIDES**

**FADED TEXT OR DRAWING**

**BLURRED OR ILLEGIBLE TEXT OR DRAWING**

**SKEWED/SLANTED IMAGES**

**COLOR OR BLACK AND WHITE PHOTOGRAPHS**

**GRAY SCALE DOCUMENTS**

**LINES OR MARKS ON ORIGINAL DOCUMENT**

**REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY**

**OTHER:** \_\_\_\_\_

## **IMAGES ARE BEST AVAILABLE COPY.**

**As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.**